STATISTICS OF EXTREMES, WITH APPLICATIONS IN ENVIRONMENT, INSURANCE AND FINANCE

Richard L. Smith

Department of Statistics
University of North Carolina
Chapel Hill, NC 27599-3260,
USA

Email address: rls@email.unc.edu

Web reference:
http://www.stat.unc.edu/postscript/rs/semstatrls.ps

12 MARCH 2003

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

The material in this paper formed the content of three lectures given as part of the SEMSTAT conference on Extreme Value Theory in Gothenburg, Sweden, December 10–15, 2001. The proceedings of the conference are to be published by CRC Press/Chapman and Hall.

I am grateful to the organisers of the conference, particularly Claudia Klüppelberg, Barbel Finkenstadt and Holger Rootzén, for inviting me to speak and for their efficient organisation of the meeting, and to the rest of the participants for their feedback during the meeting.

Part of the research reported here was carried out during the program entitled *Managing Risk*, held at the Isaac Newton Institute in Cambridge, in July–August, 2001. I am grateful to the organisers of the program, in particular Dr. Dougal Goodman of the (U.K.) Foundation for Science and Technology, and to the Director and staff of the Newton Institute for their support of this program. The Newton Institute is supported by a grant from the (U.K.) Engineering and Physical Sciences Research Council. I would also like to acknowledge the support of my personal research through the National Science Foundation, in particular grants DMS-9971980 and DMS-0084375.

1. MOTIVATING EXAMPLES

Extreme value theory is concerned with probabilistic and statistical questions related to very high or very low values in sequences of random variables and in stochastic processes. The subject has a rich mathematical theory and also a long tradition of applications in a variety of areas. Among many excellent books on the subject, Embrechts *et al.* (1997) give a comprehensive survey of the mathematical theory with an orientation towards applications in insurance and finance, while the recent book by Coles (2001) concentrates on data analysis and statistical inference for extremes.

The present survey is primarily concerned with statistical applications, and especially with how the mathematical theory can be extended to answer a variety of questions of interest to an applied scientist. Traditionally, extreme value theory has been employed to answer questions relating to the distribution of extremes (e.g., what is the probability that a windspeed over a given level will occur in a given location during a given year?) or the inverse problem of return levels (e.g. what height of a river will be exceeded with probability 1/100 in a given year? — this quantity is often called the 100-year return level). During the last 30 years, many new techniques have been developed concerned with exceedances over high thresholds, the dependence among extreme events in various types of stochastic processes, and multivariate extremes. These new techniques make it possible to answer much more complex questions than simple distributions of extremes. Among those considered in the present review are whether probabilities of extreme events are changing with time or corresponding to other measured covariates (e.g. Sections 5.1–5.3, 6.4), the simultaneous fitting of extreme value distributions to several related time series (Sections 6.1–6.3), the spatial dependence of extreme value distributions (Section 6.4) and the rather complex forms of extreme value calculations that arise in connection with financial time series (Section 8). Along the way, we shall also review relevant parts of the mathematical theory for univariate extremes (Sections 2–4) and one recent approach (among several that are available) to the characterisation of multivariate extreme value distributions (Section 7).

For the rest of this section, we give some specific examples of data-oriented questions which will serve to motivate the rest of the chapter.

1.1 Snowfall in North Carolina

On January 25, 2000, a snowfall of 20.3 inches was recorded at Raleigh-Durham airport, North Carolina. This is an exceptionally high snowfall for this part of the United States, and caused widespread disruption to travel, power supplies and the local school system. Various estimates that appeared in the press at the time indicated that such an event could be expected to occur once every 100-200 years. The question we consider here is how well one could estimate the probability of such an event based on data that were available prior to the actual event. Associated with this is the whole question of what is the uncertainty of such an assessment of an extreme value probability.

To simplify the question and to avoid having to consider time-of-year effects, we shall confine our discussion to the month of January, implicitly assuming that an extreme snowfall event is equally likely to occur at any time during the month. Thus the question we are trying to answer is, for any large value of x, "What is the probability that a snowfall exceeding x inches occurs at Raleigh-Durham airport, some time during the month of January, in any given year?"

A representative data set was compiled from the publicly available data base of the U.S. National Climatic Data Center. Table 1 lists all the January snow events (i.e. daily totals where a non-zero snowfall was recorded) at Raleigh-Durham airport, for the period 1948–1998. We shall take this as a data base from which we try to answer the question just quoted, with $x = 20.3$, for some arbitrary year after 1998. It can be seen that no snowfall anywhere close to 20.3 inches occurs in the given data set, the largest being 9.0 inches on January 19, 1955. There are earlier records of daily snowfall event over 20 inches in this region, but these were prior to the establishment of a regular series of daily measurements, and we shall not take them into account.

In Section 3, we shall return to this example and show how a simple threshold-based analysis may be used to answer this question, but with particular attention to the sensitivity to the chosen threshold and to the contrast between maximum likelihood and Bayesian approaches.

Year	Day	Amount	Year	Day	Amount	Year	Day	Amount
1948	24	1.0	1965	15	0.8	1977	7	0.3
1948	31	2.5	1965	16	3.7	1977	24	1.8
1954	11	1.2	1965	17	1.3	1979	31	0.4
1954	22	1.2	1965	30	3.8	1980	30	1.0
1954	23	4.1	1965	31	0.1	1980	31	1.2
1955	19	9.0	1966	16	0.1	1981	30	2.6
1955	23	3.0	1966	22	0.2	1982	13	1.0
1955	24	1.0	1966	25	2.0	1982	14	5.0
1955	27	1.4	1966	26	7.6	1985	20	1.7
1956	23	2.0	1966	27	0.1	1985	28	2.4
1958	7	3.0	1966	29	1.8	1987	25	0.1
1959	8	1.7	1966	30	0.5	1987	26	0.5
1959	16	1.2	1967	19	0.5	1988	7	7.1
1961	21	1.2	1968	10	0.5	1988	8	0.2
1961	26	1.1	1968	11	1.1	1995	23	0.7
1962	1	1.5	1968	25	1.4	1995	30	0.1
1962	10	5.0	1970	12	1.0	1996	6	2.7
1962	19	1.6	1970	23	1.0	1996	7	2.9
1962	28	2.0	1973	7	0.7	1997	11	0.4
1963	26	0.1	1973	8	5.7	1998	19	2.0
1964	13	0.4	1976	17	0.4			

Table 1. January snow events at Raleigh-Durham airport, 1948–1998.

1.2 Insurance Risk of a Large Company

This example is based on Smith and Goodman (2000). A data set was compiled consisting of insurance claims made by an international oil company over a 15-year period. In the data set originally received from the company, 425 claims were recorded over a nominal threshold level, expressed in U.S. dollars and adjusted for inflation to 1994 cost equivalents. As a preliminary to the detailed analysis, two further preprocessing steps were performed: (i) the data were multiplied by a common but unspecified scaling factor — this has the effect of concealing the precise sums of money involved, without in any other way changing the characteristics of the data set, (ii) simultaneous claims of the same type arising on the same day were aggregated into a single total claim for that day — the motivation for this was to avoid possible clustering effects due to claims arising from the same cause, though it is likely that this effect is minimal for the data set under consideration. With these two changes to the original data set, the analysed data consisted of 393 claims over a nominal threshold of 0.5, grouped into seven "types" as shown in Table 2.

Type	Description	Number	Mean
1	Fire	175	11.1
2	Liability	17	12.2
3	Offshore	40	9.4
4	Cargo	30	3.9
5	Hull	85	2.6
6	Onshore	44	2.7
7	Aviation	2	1.6

Table 2. The seven types of insurance claim, with the total number of claims and the mean size of claim for each type.

The total of all 393 claims was 2989.6, and the ten largest claims, in order, were 776.2, 268.0, 142.0, 131.0, 95.8, 56.8, 46.2, 45.2, 40.4, 30.7. These figures give some indication of the type of data were are talking about: the total loss to the company is dominated by the value of a few very large claims, with the largest claim itself accounting for 26% of the total. In statistical terms, the data clearly represent a very skewed, long-tailed distribution, though these features are entirely typical of insurance data.

Further information about the data can be gained from Fig. 1, which shows (a) a scatterplot of the individual claims against time — note that claims are drawn on a logarithmic scale; (b) cumulative number of claims against time — this serves as a visual indicator of whether there are trends in the frequency of claims; (c) cumulative claim amounts against time, as an indicator of trends in the total amounts of claims; (d) the so-called mean excess plot, in which for a variety of possible thresholds, the mean excess over the threshold was computed for all claims that were above that threshold, and plotted against the threshold itself. As will be seen later (Section 4), this is a useful diagnostic of the Generalised Pareto Distribution (GPD) which is widely used as a probability distribution for excesses over thresholds — in this case, the fact that the plot in Fig. 1(d) is close to a straight line over most of its range is an indicator that the GPD fits the data reasonably well. Of the other plots in Fig. 1, plot (b) shows no visual evidence of a trend in the frequency of claims, while in (c), there is a sharp rise in the cumulative total of claims during year 7, but this arises largely because the two largest claims in the whole series were both in the same year, which raises the question of whether these two claims should be treated as outliers and therefore analysed separately from the rest of the data. The case for doing this is strengthened by the fact that these were the only two claims in the entire data set that resulted from the total loss of a facility. We shall return to these issues when the data are analysed in detail in Section 6, but for the moment, we list four possible questions for discussion:

1. What is the distribution of very large claims?
2. Is there any evidence of a change of the distribution of claim sizes and frequencies over time?
3. What is the influence of the different types of claim on the distribution of total claim size?
4. How should one characterise the risk to the company? More precisely, what probability distribution can one put on the amount of money that the company will have to pay out in settlement of large insurance claims over a future time period of, say, one year?

Published statistical analyses of insurance data often concentrate exclusively on question 1, but it is arguable that the other three questions are all more important and relevant than a simple characterisation of the probability distribution of claims, for a company planning its future insurance policies.

1.3 Value at Risk in Finance

Much of the recent research in extreme value theory has been stimulated by the possibility of large losses in the financial markets, which has resulted in a large literature on "Value at Risk" and other measures of financial vulnerability. As an example of the types of data analysed and the kinds of questions asked, Fig. 2 shows negative daily returns from closing prices of 1982-2001 stock prices in three companies, Pfizer, General Electric and Citibank. If X_t is the closing price of a stock or financial index on day t, then the daily return (in effect, the percentage loss or gain on the day) is defined either by

$$Y_t = 100\left(\frac{X_t}{X_{t-1}} - 1\right) \tag{1.1}$$

or, more conveniently for the present discussion, by

$$Y_t = 100\ \log\frac{X_t}{X_{t-1}}. \tag{1.2}$$

We are mainly interested in the possibility of large losses rather than large gains, so we rewrite (1.2) in terms of negative returns,

$$Y_t = 100\ \log\frac{X_{t-1}}{X_t}, \tag{1.3}$$

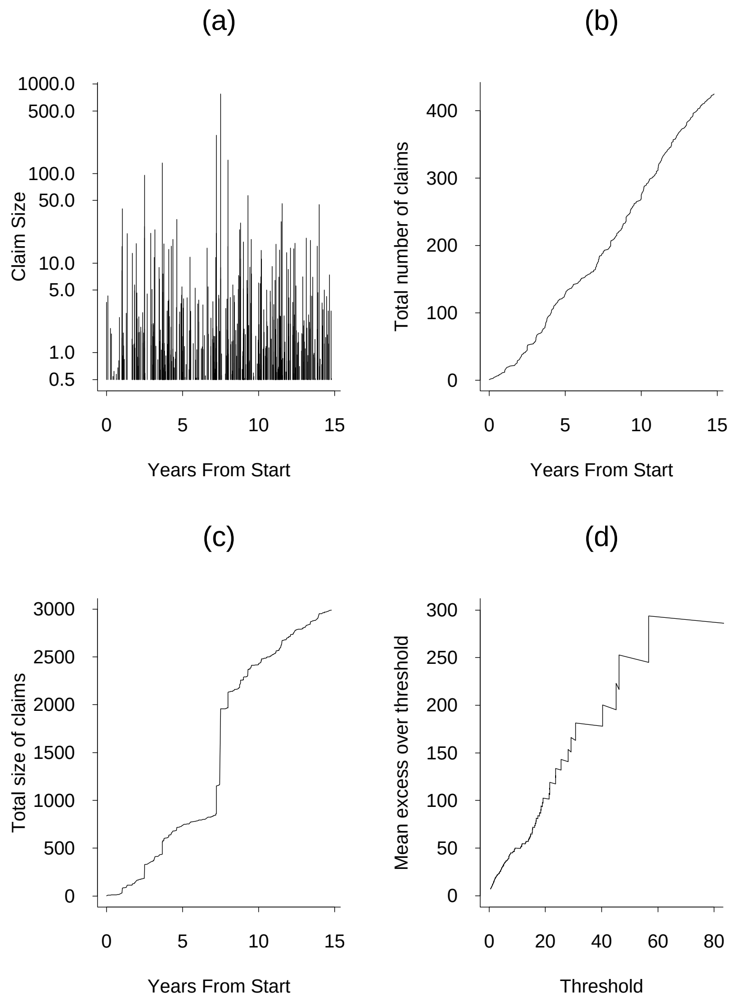


Figure 1. Insurance data. (a) Plot of raw data. (b) Cumulative number of claims *vs.* time. (c) Cumulative claim amount *vs.* time. (d) Mean excess plot.

which is the quantity actually plotted in Fig. 2.

Typical problems here are:

1. Calculating the Value at Risk, i.e. the amount which might be lost in a portfolio of assets over a specified time period with a specified small probability;
2. Describing dependence among the extremes of different series, and using this description in the problem of managing a portfolio of investments;
3. Modeling extremes in the presence of volatility — like all financial time series, those in Fig. 2 show periods when the variability or volatility in the series is high, and others where it is much lower, but simple theories of extreme values in independent and identically distributed (i.i.d.) random variables or simple stationary time series do not account for such behaviour.

In Section 8, we shall return to this example and suggest some possible approaches to answering these questions.

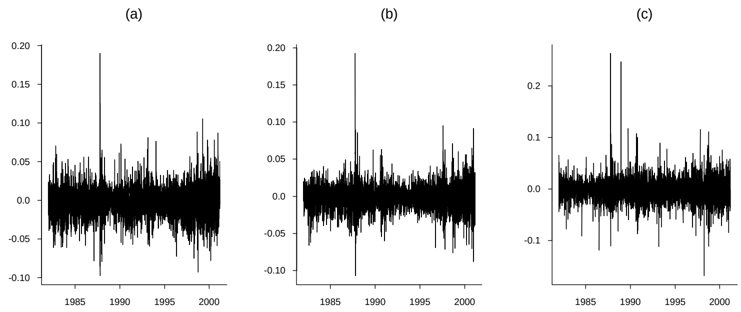


Figure 2. Negative daily returns, defined by (1.3), for three stocks, 1982–2001. (a) Pfizer, (b) General Electric, (c) Citibank.

2. UNIVARIATE EXTREME VALUE THEORY

2.1 The Extreme Value Distributions

In this section, we outline the basic theory that applies to univariate sequences of i.i.d. random variables. This theory is by now very well established, and is the starting point for all the extreme value methods we shall discuss.

Suppose we have an i.i.d. sequence of random variables, $X_1, X_2, ...$, whose common cumulative distribution function is F, i.e.

$$F(x) = \Pr\{X_i \le x\}.$$

Also let $M_n = \max(X_1, ..., X_n)$ denote the nth sample maximum of the process. Then

$$\Pr\{M_n \le x\} = F(x)^n. \tag{2.1}$$

Result (2.1) is of no immediate interest, since it simply says that for any fixed x for which $F(x) < 1$, we have $\Pr\{M_n \le x\} \to 0$. For non-trivial limit results we must *renormalise*: find $a_n > 0, b_n$ such that

$$\Pr\left\{\frac{M_n - b_n}{a_n} \le x\right\} = F(a_n x + b_n)^n \to H(x). \tag{2.2}$$

The *Three Types Theorem*, originally stated without detailed mathematical proof by Fisher and Tippett (1928), and later derived rigorously by Gnedenko (1943), asserts that if a nondegenerate H exists (i.e. a distribution function which does not put all its mass at a single point), it must be one of three types:

$$H(x) = \exp(-e^{-x}), \text{ all } x, \tag{2.3}$$

$$H(x) = \begin{cases} 0, & x < 0, \\ \exp(-x^{-\alpha}), & x > 0, \end{cases} \tag{2.4}$$

$$H(x) = \begin{cases} \exp(-|x|^{\alpha}), & x < 0, \\ 1, & x > 0, \end{cases} \tag{2.5}$$

Here two distribution functions H_1 and H_2 are said to be of the same type if one can be derived from the other through a simple location-scale transformation,

$$H_1(x) = H_2(Ax + B), \quad A > 0.$$

Very often, (2.3) is called the *Gumbel type*, (2.4) the *Fréchet type* and (2.5) the *Weibull type*. In (2.4) and (2.5), $\alpha > 0$.

The three types may be combined into a single *Generalised Extreme Value* (GEV) distribution:

$$H(x) = \exp\left\{-\left(1+\xi\frac{x-\mu}{\psi}\right)_+^{-1/\xi}\right\}, \tag{2.6}$$

($y_+ = \max(y, 0)$) where μ is a location parameter, $\psi > 0$ is a scale parameter and ξ is a shape parameter. The limit $\xi \to 0$ corresponds to the Gumbel distribution, $\xi > 0$ to the Fréchet distribution with $\alpha = 1/\xi$, $\xi < 0$ to the Weibull distribution with $\alpha = -1/\xi$.

In more informal language, the case $\xi > 0$ is the "long-tailed" case for which $1 - H(x) \propto x^{-1/\xi}$, $\xi = 0$ is the "medium-tailed" case for which $1 - H(x)$ decreases exponentially for large x, and $\xi < 0$ is the "short-tailed" case, in which the distribution has a finite endpoint (the minimum value of x for which $H(x) = 1$) at $x = \mu - \psi/\xi$.

2.2 Exceedances Over Thresholds

Consider the distribution of X conditionally on exceeding some high threshold u (so $Y = X - u > 0$):

$$F_u(y) = \Pr\{Y \le y \mid Y > 0\} = \frac{F(u+y) - F(u)}{1 - F(u)}.$$

As $u \to \omega_F = \sup\{x : \; F(x) < 1\}$, we often find a limit

$$F_u(y) \approx G(y; \sigma_u, \xi), \tag{2.7}$$

where G is *Generalised Pareto Distribution* (GPD)

$$G(y; \sigma, \xi) = 1 - \left(1 + \xi\frac{y}{\sigma}\right)_+^{-1/\xi}. \tag{2.8}$$

Although the Pareto and similar distributions have long been used as models for long-tailed processes, the rigorous connection with classical extreme value theory was established by Pickands (1975). In effect, Pickands showed that for any given F, a GPD approximation arises from (2.7) if and only there exist normalising constants and a limiting H such that the classical extreme value limit result (2.2) holds; in that case, if H is written in GEV form (2.6), then the shape parameter ξ is the same as the corresponding GPD parameter in (2.8). Thus there is a close parallel between limit results for sample maxima and limit results for exceedances over thresholds, which is quite extensively exploited in modern statistical methods for extremes.

In the GPD, the case $\xi > 0$ is long-tailed, for which $1 - G(x)$ decays at the same rate as $x^{-1/\xi}$ for large x. This is reminiscent of the usual Pareto distribution, $G(x) = 1 - cx^{-\alpha}$, with $\xi = 1/\alpha$. For $\xi = 0$, we may take the limit as $\xi \to 0$ to get

$$G(y; \sigma, 0) = 1 - \exp\left(-\frac{y}{\sigma}\right),$$

i.e. exponential distribution with mean σ. For $\xi < 0$, the distribution has finite upper endpoint at $-\sigma/\xi$. Some other elementary results about the GPD are

$$\begin{aligned} \mathrm{E}(Y) &= \frac{\sigma}{1-\xi}, \;(\xi < 1), \\ \mathrm{Var}(Y) &= \frac{\sigma^2}{(1-\xi)^2(1-2\xi)}, \;(\xi < \frac{1}{2}), \\ \mathrm{E}(Y - y \mid Y > y > 0) &= \frac{\sigma + \xi y}{1-\xi}, \;(\xi < 1). \end{aligned} \tag{2.9}$$

2.2.1 Poisson-GPD model for exceedances

Suppose we observe i.i.d. random variables $X_1, ..., X_n$, and observe the indices i for which $X_i > u$. If these indices are rescaled to points i/n, they can be viewed as a point process of rescaled exceedance times on $[0, 1]$. If $n \to \infty$ and $1 - F(u) \to 0$ such that $n(1 - F(u)) \to \lambda$ $(0 < \lambda < \infty)$, the process converges weakly to a homogeneous Poisson process on $[0, 1]$, of intensity λ.

Motivated by this, we can imagine a limiting form of the joint point process of exceedance times and excesses over the threshold, of the following form:

(a) The number, N, of exceedances of the level u in any one year has a Poisson distribution with mean λ;

(b) Conditionally on $N \geq 1$, the excess values $Y_1, ..., Y_N$ are i.i.d. from the GPD.

We call this the *Poisson-GPD model*.

Of course, there is nothing special here about one year as the unit of time — we could just as well use any other time unit — but for environmental processes in particular, a year is often the most convenient reference time period.

The Poisson-GPD process is closely related to the GEV distribution for annual maxima. Suppose $x > u$. The probability that the annual maximum of the Poisson-GPD process is less than x is

$$\begin{aligned} \Pr\{\max_{1 \leq i \leq N} Y_i \leq x\} &= \Pr\{N = 0\} + \sum_{n=1}^{\infty} \Pr\{N = n,\ Y_1 \leq x,\ ...\ Y_n \leq x\} \\ &= e^{-\lambda} + \sum_{n=1}^{\infty} \frac{\lambda^n e^{-\lambda}}{n!} \cdot \left\{1 - \left(1 + \xi \frac{x - u}{\sigma}\right)_+^{-1/\xi}\right\}^n \\ &= \exp\left\{-\lambda \left(1 + \xi \frac{x - u}{\sigma}\right)_+^{-1/\xi}\right\}. \end{aligned} \tag{2.10}$$

If we substitute

$$\sigma = \psi + \xi(u - \mu), \qquad \lambda = \left(1 + \xi \frac{u - \mu}{\psi}\right)^{-1/\xi}, \tag{2.11}$$

(2.10) reduces to the GEV form (2.6). Thus the GEV and GPD models are entirely consistent with one another above the threshold u, and (2.11) gives an explicit relationship between the two sets of parameters.

The Poisson-GPD model is closely related to the *Peaks Over Threshold* (POT) model originally developed by hydrologists. In cases with high serial correlation the threshold exceedances do not occur singly but in clusters, and in that case, the method is most directly applied to the peak values within each cluster. For more detailed discussion see Davison and Smith (1990).

Another issue is *seasonal dependence*. For environmental processes in particular, it is rarely the case that the probability of an extreme event is independent of the time of year, so we need some extension of the model to account for seasonality. Possible strategies include:

(a) Remove seasonal trend before applying the threshold approach;

(b) Apply the Poisson–GPD model separately to each season;

(c) Expand the Poisson–GPD model to include covariates.

All three approaches have been extensively applied in past discussions of threshold methods. In the present chapter, we shall focus primarily on method (c) (e.g. Sections 5 and 6.4), though only after first rewriting the Poisson-GPD model in a different form (Section 2.5).

2.3 Examples

In this section, we present four examples to illustrate how the extreme value and GPD limiting distributions work in practice, given various assumptions on the distribution function F from which the random

variables are drawn. From a mathematical viewpoint, these examples are all special cases of the *domain of attraction* problem, which has been dealt with extensively in texts on extreme value theory, e.g. Leadbetter *et al.* (1983) or Resnick (1987). Here we make no attempt to present the general theory, but the examples serve to illustrate the concepts in some of the most typical cases.

2.3.1 The exponential distribution.

Suppose $F(x) = 1 - e^{-x}$. Let $a_n = 1,\ b_n = \log n$. Then

$$\begin{aligned} F^n(a_n x + b_n) &= \left(1 - e^{-x-\log n}\right)^n \\ &= \left(1 - \frac{e^{-x}}{n}\right)^n \\ &\to \exp\left(-e^{-x}\right), \end{aligned}$$

in other words, the limiting distribution of sample extremes in this case is the Gumbel distribution.

For the threshold version of the result, set $\sigma_u = 1$. Then

$$\begin{aligned} F_u(\sigma_u z) &= \frac{F(u+z) - F(u)}{1 - F(u)} \\ &= \frac{e^{-u} - e^{-u-z}}{e^{-u}} \\ &= 1 - e^{-z} \end{aligned}$$

so the exponential distribution of mean 1 is the exact distribution for exceedances in this case.

2.3.2 Pareto-type tail

Suppose $1 - F(x) \sim cx^{-\alpha}$ as $x \to \infty$, with $c > 0$, $\alpha > 0$. Let $b_n = 0,\ a_n = (nc)^{1/\alpha}$. Then for $x > 0$,

$$\begin{aligned} F^n(a_n x) &\approx \left\{1 - c\left(a_n x\right)^{-\alpha}\right\}^n \\ &= \left(1 - \frac{x^{-\alpha}}{n}\right)^n \\ &\to \exp\left(-x^{-\alpha}\right), \end{aligned}$$

which is the Fréchet limit.

For the threshold result, let $\sigma_u = ub$ for some $b > 0$. Then

$$\begin{aligned} F_u(\sigma_u z) &= \frac{F(u+ubz) - F(u)}{1 - F(u)} \\ &\approx \frac{cu^{-\alpha} - c(u+ubz)^{-\alpha}}{cu^{-\alpha}} \\ &= 1 - (1+bz)^{-\alpha}. \end{aligned}$$

Set $\xi = \frac{1}{\alpha},\ b = \xi$ to get the result in GPD form.

2.3.3 Finite upper endpoint

Suppose $\omega_F = \omega < \infty$ and $1 - F(\omega - y) \sim cy^\alpha$ as $y \downarrow 0$ for $c > 0, \alpha > 0$. Set $b_n = \omega, a_n = (nc)^{-1/\alpha}$. Then for $x < 0$,

$$\begin{aligned} F^n(a_n x + b_n) &= F^n(\omega + a_n x) \\ &\approx \{1 - c\,(-a_n x)^\alpha\}^n \\ &= \left\{1 - \frac{(-x)^\alpha}{n}\right\}^n \\ &\to \exp\{-(-x)^\alpha\}, \end{aligned}$$

which is of Weibull type.

For the threshold version, let u be very close to ω and consider $\sigma_u = b(\omega - u)$ for $b > 0$ to be determined. Then for $0 < z < \frac{1}{b}$,

$$\begin{aligned} F_u(\sigma_u z) &= \frac{F(u+\sigma_u z) - F(u)}{1 - F(u)} \\ &\approx \frac{c(\omega-u)^\alpha - c(\omega - u - \sigma_u z)^\alpha}{c(\omega-u)^\alpha} \\ &= 1 - (1 - bz)^\alpha. \end{aligned}$$

This is of GPD form with $\xi = -\frac{1}{\alpha}$ and $b = -\xi$.

2.3.4 Normal extremes

Let $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2} dy$. By Feller (1968), page 193,

$$1 - \Phi(x) \sim \frac{1}{x\sqrt{2\pi}} e^{-x^2/2} \quad \text{as } x \to \infty.$$

Then

$$\begin{aligned} \lim_{u\to\infty} \frac{1 - \Phi(u + z/u)}{1 - \Phi(u)} &= \lim_{u\to\infty} \left[\left(1 + \frac{z}{u^2}\right)^{-1} \cdot \exp\left\{ -\frac{1}{2}\left(u + \frac{z}{u}\right)^2 + \frac{1}{2}u^2 \right\} \right] \\ &= e^{-z}. \end{aligned}$$

For a first application, let $\sigma_u = 1/u$, then

$$\frac{\Phi(u+\sigma_u z) - \Phi(u)}{1 - \Phi(u)} \to 1 - e^{-z} \quad \text{as } u \to \infty,$$

so the limiting distribution of exceedances over thresholds is exponential.

For a second application, define b_n by $\Phi(b_n) = 1 - 1/n, a_n = 1/b_n$. Then

$$\begin{aligned} n\left\{1 - \Phi\left(a_n x + b_n\right)\right\} &= \frac{1 - \Phi\left(a_n x + b_n\right)}{1 - \Phi\left(b_n\right)} \\ &\to e^{-x} \end{aligned}$$

and hence

$$\lim_{n\to\infty} \Phi^n\left(a_n x + b_n\right) = \lim_{n\to\infty} \left(1 - \frac{e^{-x}}{n}\right)^n = \exp\left(-e^{-x}\right),$$

establishing convergence to Gumbel limit.

In practice, although the Gumbel and exponential distributions are the correct limits for sample maxima and threshold exceedances respectively, in practice better approximations are obtained using the GEV and GPD, allowing $\xi \neq 0$. This is known as the *penultimate approximation* and was investigated in detail by Cohen (1982a, 1982b). The practical implication of this is that it is generally better to use the GEV/GPD distributions even when we suspect that Gumbel/exponential are the correct limits.

2.4 The r Largest Order Statistics Model

An extension of the annual maximum approach is to use the r largest observations in each fixed time period (say, one year), where $r > 1$. The mathematical result on which this relies is that (2.2) is easily extended to the joint distribution of the r largest order statistics, as $n \to \infty$ for a fixed $r > 1$, and this may therefore be used as a basis for statistical inference. A practical caution is that the r-largest result is more vulnerable to departures from the i.i.d. assumption (say, if there is seasonal variation in the distribution of observations, or if observations are dependent) than the classical results about extremes.

The main result is as follows: if $Y_{n,1} \geq Y_{n,2} \geq ... \geq Y_{n,r}$ are r largest order statistics of i.i.d. sample of size n, and a_n and b_n are the normalising constants in (2.2), then

$$\left(\frac{Y_{n,1} - b_n}{a_n}, ..., \frac{Y_{n,r} - b_n}{a_n}\right)$$

converges in distribution to a limiting random vector $(X_1, ..., X_r)$, whose density is

$$h(x_1, ..., x_r) = \psi^{-r} \exp\left\{ - \left(1 + \xi \frac{x_r - \mu}{\psi}\right)^{-1/\xi} - \left(1 + \frac{1}{\xi}\right) \sum_{j=1}^{r} \log\left(1 + \xi \frac{x_j - \mu}{\psi}\right)\right\}. \quad (2.12)$$

Some examples using this approach are the papers of Smith (1986) and Tawn (1988) on hydrological extremes, and Robinson and Tawn (1995) and Smith (1997) for a novel application to the analysis of athletic records. The latter application is discussed in Section 3.3.

2.5 Point Process Approach

This was introduced as a statistical approach by Smith (1989), though the basic probability theory from which it derives had been developed by a number of earlier authors. In particular, the books by Leadbetter *et al.* (1983) and Resnick (1987) contain much information on point-process viewpoints of extreme value theory.

In this approach, instead of considering the times at which high-threshold exceedances occur and the excess values over the threshold as two separate processes, they are combined into one process based on a two-dimensional plot of exceedance times and exceedance values. The asymptotic theory of threshold exceedances shows that under suitable normalisation, this process behaves like a *nonhomogeneous Poisson process.*

In general, a nonhomogeneous Poisson process on a domain $\mathcal{D}$ is defined by an intensity $\lambda(x),\ x \in \mathcal{D}$, such that if A is a measurable subset of $\mathcal{D}$ and $N(A)$ denotes the number of points in A, then $N(A)$ has a Poisson distribution with mean

$$\Lambda(A) = \int_A \lambda(x) dx.$$

If $A_1, A_2, ...,$ are *disjoint* subsets of $\mathcal{D}$, then $N(A_1), N(A_2), ...$ are independent Poisson random variables.

For the present application, we assume x is two-dimensional and identified with (t, y) where t is time and $y \geq u$ is the value of the process, $\mathcal{D} = [0, T] \times [u, \infty)$, and we write

$$\lambda(t, y) = \frac{1}{\psi}\left(1 + \xi \frac{y - \mu}{\psi}\right)^{-1/\xi - 1}, \quad (2.13)$$

defined wherever $\{1 + \xi(y - \mu)/\psi\} > 0$ (elsewhere $\lambda(t, y) = 0$). If A is a set of the form $[t_1, t_2] \times [y, \infty)$ (see Fig. 3), then

$$\Lambda(A) = (t_2 - t_1)\left(1 + \xi \frac{y - \mu}{\psi}\right)^{-1/\xi} \quad \text{provided } y \geq u,\ 1 + \xi(y - \mu)/\psi > 0. \quad (2.14)$$

The mathematical justification for this approach lies in limit theorems as $T \to \infty$ and $1 - F(u) \to 1$, which we shall not go into here. To fit the model, we note that if a nonhomogeneous process of intensity $\lambda(t, y)$ is observed on a domain $\mathcal{D}$, and if $(T_1, Y_1), ..., (T_N, Y_N)$ are the N observed points of the process, then the joint density is

$$\prod_{i=1}^{N} \lambda(T_i, Y_i) \cdot \exp\left\{ - \int_{\mathcal{D}} \lambda(t, y) dt dy \right\}, \quad (2.15)$$

so (2.15) may be treated as a likelihood function and maximised with respect to the unknown parameters of the process. In practice, the integral in (2.15) is approximated by a sum, e.g. over all days if the observations are recorded daily.

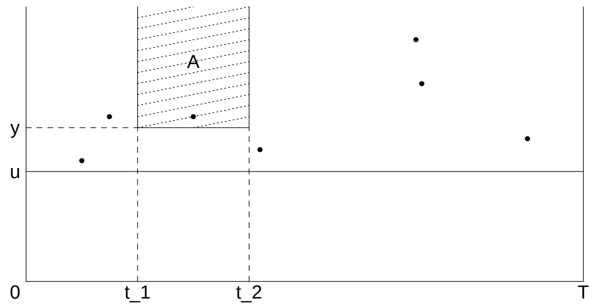


Figure 3. Illustration of point process approach. Assume the process is observed over a time interval $[0, T]$, and that all observations above a threshold level u are recorded. These points are marked on a two-dimensional scatterplot as shown in the diagram. For a set A of the form shown in the figure, the count $N(A)$ of observations in the set A is assumed to be Poisson with mean of the form given by (2.14).

An extension of this approach allows for nonstationary processes in which the parameters μ, ψ and ξ are all allowed to be time-dependent, denoted μ_t, ψ_t and ξ_t. Thus, (2.13) is replaced by

$$\lambda(t, y) = \frac{1}{\psi_t} \left(1 + \xi_t \frac{y - \mu_t}{\psi_t} \right)^{-1/\xi_t - 1}. \tag{2.16}$$

In the homogenous case where μ, ψ and ξ are constants, the model is mathematically equivalent to the Poisson-GPD model discussed in Section 2.2.1, though with a different parameterisation. The extension (2.16) is particularly valuable in connection with extreme value regression problems, which are extensively discussed later on (Sections 5 and 6).

As an illustration of how the point process viewpoint may be used as a practical guide to visualising extreme value data, Fig. 4 presents two plots derived from a 35-year series of the River Nidd in northern England (Davison and Smith 1990). The data in this case consist of daily river flows above the level of 65 cumecs/second, and have been crudely declustered to remove successive high values that are part of the same extreme event. Fig. 4(a) shows the data plotted against time measured as day within the year (1–366). Fig. 4(b) shows a corresponding plot where time is the total cumulative number of days since the start of the series in 1934. Plot (a) is a visual diagnostic for seasonality in the series and shows, not surprisingly, that there are very few exceedances during the summer months. Plot (b) may be used as a diagnostic for overall trends in the series; in this case there are three large values at the right-hand end of the series which could possibly indicate a trend in the extreme values of the series.

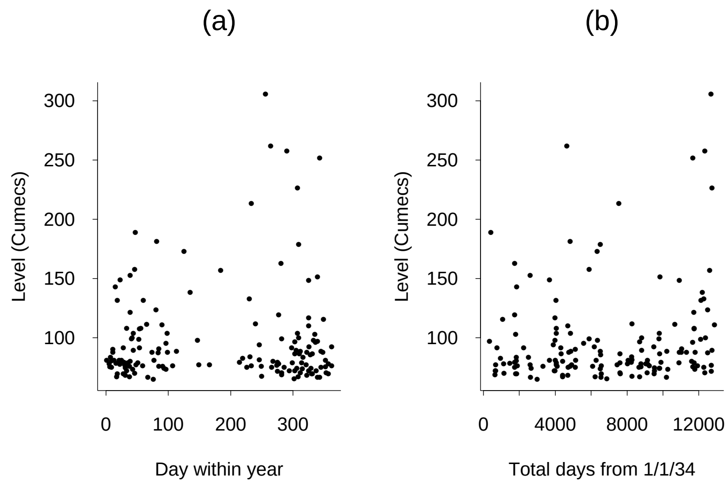


Figure 4. Plots of exceedances of River Nidd, (a) against day within year, (b) against total days from January 1, 1934. Adapted from Davison and Smith (1990).

3. ESTIMATION

3.1 Maximum Likelihood Estimation

Suppose we have observations $Y_1, ..., Y_N$ which are data for which the GEV distribution (2.6) is appropriate. For example, perhaps we take one year as the unit of time, and $Y_1, ..., Y_N$ represent annual maximum values for each of N years. The corresponding log likelihood is

$$\ell_Y(\mu, \psi, \xi) = -N \log \psi - \left(\frac{1}{\xi} + 1 \right) \sum_i \log \left(1 + \xi \frac{Y_i - \mu}{\psi} \right) - \sum_i \left(1 + \xi \frac{Y_i - \mu}{\psi} \right)^{-1/\xi} \tag{3.1}$$

provided $1 + \xi(Y_i - \mu)/\psi > 0$ for each i.

For the Poisson-GPD model of Section 2.2.1, suppose we have a total of N observations above a threshold in time T, and the excesses over the threshold are $Y_1, ..., Y_N$. Suppose the expected number of Poisson exceedances is λT, and the GPD parameters are σ and ξ, as in (2.8). Then the log likelihood is

$$\ell_{N,Y}(\lambda, \sigma, \xi) = N \log \lambda - \lambda T - N \log \sigma - \left(1 + \frac{1}{\xi} \right) \sum_{i=1}^{N} \log \left(1 + \xi \frac{Y_i}{\sigma} \right) \tag{3.2}$$

provided $1 + \xi Y_i/\sigma > 0$ for all i.

Similar log likelihoods may be constructed from the joint densities (2.12) and (2.15) for the r largest order statistics approach and the point process approach.

The maximum likelihood estimators are the values of the unknown parameters that maximise the log likelihood. In practice these are are local maxima found by nonlinear optimisation. The standard asymptotic results of consistency, asymptotic efficiency and asymptotic normality hold for these distributions if $\xi > -\frac{1}{2}$ (Smith 1985). In particular, the elements of the Hessian matrix of $-\ell$ (the matrix of second-order partial derivatives, evaluated at the maximum likelihood estimators) are known as the *observed information matrix*, and the inverse of this matrix is a widely used approximation for the variance-covariance matrix of the maximum likelihood estimators. The square roots of the diagonal entries of this inverse matrix are estimates of the standard deviations of the three parameter estimates, widely known as the *standard errors* of those estimates. All these results are asymptotic approximations valid for large sample sizes, but in practice they are widely used even when the sample sizes are fairly small.

3.2 Profile Likelihoods for Quantiles

Suppose we are interested in the n-year return level y_n, i.e. the $(1-1/n)$-quantile of the annual maximum distribution. This is given by solving the equation

$$\exp \left\{ - \left(1 + \xi \frac{y_n - \mu}{\psi} \right)^{-1/\xi} \right\} = 1 - \frac{1}{n}. \tag{3.2}$$

Exploiting the approximation $1 - \frac{1}{n} \approx \exp\left(-\frac{1}{n}\right)$, this simplifies to

$$\left(1 + \xi \frac{y_n - \mu}{\psi} \right)^{-1/\xi} = \frac{1}{n}$$

and hence

$$y_n = \mu + \psi \frac{n^\xi - 1}{\xi}. \tag{3.3}$$

One approach to the estimation of y_n is simply to substitute the maximum likelihood estimates $\hat{\mu}, \hat{\psi}, \hat{\xi}$ for the unknown parameters μ, ψ, ξ, thus creating a maximum likelihood estimator $\hat{y}_n$. The variance of $\hat{y}_n$

may be estimated through a standard delta function approximation, i.e. if we define a vector of partial derivatives

$$g(\mu, \psi, \xi) = \left(\frac{\partial y_n}{\partial \mu}, \frac{\partial y_n}{\partial \psi}, \frac{\partial y_n}{\partial \xi} \right)$$

and write $\hat{g}$ for $g(\hat{\mu}, \hat{\psi}, \hat{\xi})$, and also write H for the observed information matrix for $(\hat{\mu}, \hat{\psi}, \hat{\xi})$, then the variance of $\hat{y}_n$ is approximately

$$\hat{g} \cdot H^{-1} \cdot \hat{g}^T, \tag{3.4}$$

and the square root of (3.4) is an approximate standard error. In practice, this often gives a rather poor approximation which does not account for the skewness of the distribution of $\hat{y}_n$, especially when n is large.

An alternative approach is via a profile likelihood. Equation (3.3) shows how y_n (for a given value of n) may be expressed as a function of (μ, ψ, ξ). Suppose we rewrite this as

$$\mu = y_n - \psi \frac{n^{\xi} - 1}{\xi}$$

and substitute in (3.1) so that the log likelihood ℓ_Y is written as a function of new parameters (y_n, ψ, ξ). If for any given value of y_n we maximise this function with respect to ψ and ξ, we obtain a function of y_n alone, say $\ell_Y^*(y_n)$, which is known as the *profile log likelihood* function for y_n. This function may be plotted to determine the relative plausibility of different values of y_n.

A confidence interval for y_n may be constructed by exploiting the following property: under standard regularity conditions for maximum likelihood (which, as noted already, are valid provided $\xi > -\frac{1}{2}$), under the null hypothesis that y_n is the true value,

$$2\{\log \ell_Y^*(\hat{y}_n) - \log \ell_Y^*(y_n)\} \sim \chi_1^2 \text{ (approximately)}. \tag{3.5}$$

Therefore, an approximate $100(1-\alpha)\%$ confidence interval for y_n consists of all values for which

$$\log \ell_Y^*(\hat{y}_n) - \log \ell_Y^*(y_n) \leq \frac{1}{2} \chi^2_{1;1-\alpha} \tag{3.5}$$

where $\chi^2_{1;1-\alpha}$ is the $(1-\alpha)$-quantile of the χ_1^2 distribution. For example, in the case $\alpha = .05$, the right-hand side of (3.5) is 1.92.

The same concept may be used in connection with (3.2) or any other model for which the standard regularity conditions for maximum likelihood hold. For example, Fig. 5 (adapted from Davison and Smith, 1990) shows a profile likelihood plot for y_n, for each of the values $n = 25$, 50 and 100, constructed from the River Nidd threshold exceedances mentioned already in Section 2.4. The key point here is that the curves are highly skewed to the right, and correspondingly, so are the confidence intervals — in sharp contrast to the confidence intervals derived from the delta method that are always symmetric about the maximum likelihood estimator. This in turn reflects that there is much less information in the data about the behaviour of the process at very high threshold levels (above 400 say) compared with lower levels where there is much more data. Although there is no proof that the confidence intervals derived by the profile likelihood method necessarily have better coverage probabilities than those derived by the delta method, simulations and practical experience suggest that they do.

3.3 Bayesian Approaches

Bayesian methods of statistics are based on specifying a density function for the unknown parameters, known as the prior density, and then computing a posterior density for the parameters given the observations. In practice, such computations are nearly always carried out using some form of Markov chain Monte Carlo (MCMC) sampling, which we shall not describe here as a number of excellent texts on the subject are available, e.g. Gamerman (1997) or Robert and Casella (2000). In the present discussion, we shall not dwell on the philosophical differences between Bayesian and frequentist approaches to statistics, but concentrate on two features that may be said to give Bayesian methods a practical advantage: their effectiveness in

handling models with very large numbers of parameters (in particular, *hierarchical models*, which we shall see in an extreme values context in Section 6.3); and their usefulness in *predictive inference*, where the ultimate objective is not so much to learn the values of unknown parameters but to establish a meaningful probability distribution for future unobserved random quantities.

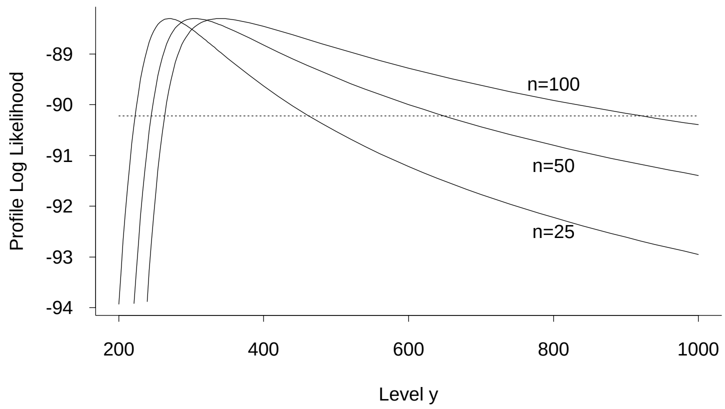


Figure 5. Profile log likelihood plots for the n-year return value y_n for the River Nidd, for n=25, 50 and 100. The horizontal dotted line is at a level 1.92 below the common maximum of the three curves; for each n, a 95% confidence interval for y_n consists of all values for which the profile log likelihood is above the dotted line.

For the rest of the present section, we focus on a specific example, first given by Smith (1997), that brings out the contrast between maximum likelihood inference about parameters and Bayesian predictive inference in a particularly striking way. Further instances of Bayesian predictive inference applied to extremes will be found in several subsequent sections, e.g. 3.4 and 6.2.

The example concerns the remarkable series of track performances achieved during 1993 by the Chinese athlete Wang Junxia, including new world records at 3,000 metres and 10,000 metres which were such an improvement on previous performances that there were immediate suspicions that they were drug assisted. However, although other Chinese athletes have tested positive for drugs, Wang herself never did, and her records still stand. The question considered here, and in an earlier paper by Robinson and Tawn (1995), is to assess just how much of an outlier the performance really was, in comparison with previous performances. The detailed discussion is confined to the 3,000 metres event.

Fig. 6(a) shows the five best running times by different athletes in the women's 3000 metre track event for each year from 1972 to 1992, along with Wang Junxia's world record from 1993. The first step is to fit a probability model to the data up to 1992.

Recall from (2.12) that there is an asymptotic distribution for the joint distribution of the r largest order statistics in a random sample, in terms of the usual GEV parameters (μ, ψ, ξ). Recall also that the upper endpoint of the distribution is at $\mu - \psi/\xi$ when $\xi < 0$. In this example, this is applied with $r = 5$, the observations (running times) are negated to convert minima into maxima, and the endpoint parameter is denoted x_{ult}, the nominal "ultimate running performance". A profile likelihood for x_{ult} may therefore be constructed by the same method as in Section 3.2. For the present study, the analysis is confined to the data from 1980 onwards, for which there is no visible evidence of a time trend.

The profile log likelihood constructed by this process is shown in Fig. 6(b). A 95% confidence interval for x_{ult} is again calculated as the set of values for which the profile log likelihood is above the dashed line, and leads to an approximate confidence interval (481,502). Wang's 1993 record — 486.1 seconds — lies within this confidence interval, so on the basis of the analysis so far, there is no clear-cut basis on which to say her record was anomalous. Robinson and Tawn (1995) considered a number of other models for the data, for example allowing for various forms of time trend from 1972 onwards, but their main conclusions were

consistent with this, i.e. that likelihood ratio confidence intervals for x_{ult} were wide and typically included Wang's record.

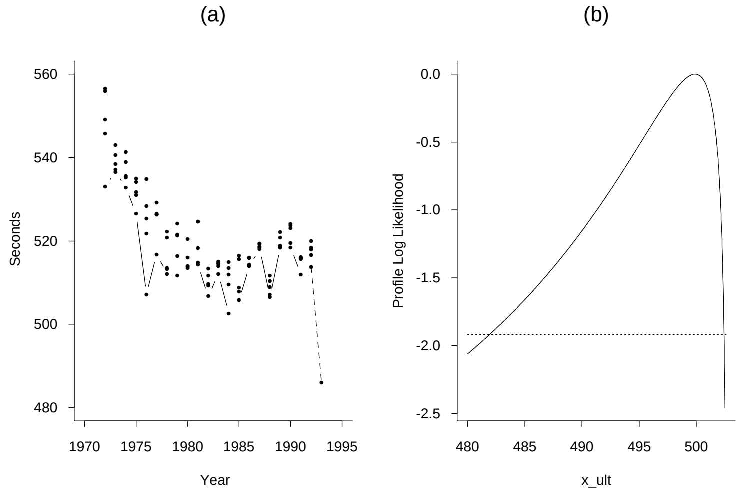


Figure 6. (a) Five best performances by different athletes in the women's 3,000 metre event, for each year from 1972 to 1992, together with Wang Junxia's record from 1993. (b) Profile log likelihood for x_{ult}, the ultimate endpoint parameter, based on the joint density (2.12) fitted to the data (multiplied by –1 to convert minima into maxima) from 1980 to 1992.

The alternative Bayesian analysis introduced by Smith (1997) was to consider the problem, not as one about estimating an ultimate limit parameter, but a more specific problem of *predicting* the best performance of 1993 given the preceding results for 1972–1992. The idea underlying this is that a prediction interval should give more precise information about what is likely to happen in a given year than the estimation of a parameter such as x_{ult}.

More precisely, Smith considered the conditional probability of a record equal to or better than the one actually achieved by Wang, given the event that the previous world record was broken. The conditioning was meant to provide some correction for the obvious selection effect, i.e. we would not even be considering these questions unless we had already observed some remarkable performance, such as a new world record. This conditional probability may be expressed as a specific analytic function of the three parameters μ, ψ and ξ, say $\phi(\mu, \psi, \xi)$, and hence estimated through the Bayesian formula

$$\int\int\int \phi(\mu, \psi, \xi)\pi(\mu, \psi, \xi|Y)d\mu d\psi d\xi \tag{3.6}$$

where $\pi(...|Y)$ denotes the posterior density given past data Y. Once again, the analysis was confined to the years 1980–1992 and did not take account of any time trend. A diffuse but proper prior distribution was assumed. The result, in this case, was .0006 (a modification of the result .00047 that was actually quoted in the paper by Smith (1997)). Such a small estimated probability provides strong evidence that Wang's performance represented an actual change in the distribution of running times. It does not, of course, provide any direct evidence that drugs were involved.

In this case, the sharp contrast between the maximum likelihood and Bayesian approaches is not a consequence of the prior distribution, nor of the MCMC method of computation, though the precise numerical result is sensitive to these. The main reason for the contrasting results lies in the change of emphasis from estimating a parameter of the model — for which the information in the data is rather diffuse, resulting in wide confidence intervals — to predicting a specific quantity, for which much more precise information is available. Note that the alternative "plug-in" approach to (3.6), in which $\phi(\mu, \psi, \xi)$ is estimated by $\phi(\hat{\mu}, \hat{\psi}, \hat{\xi})$, where $\hat{\mu}, \hat{\psi}, \hat{\xi}$ are the maximum likelihood estimators, would result in a predicted probability (of a performance as good as Wang's) of 0. This is a consequence of the fact that the maximum likelihood

estimate of x_{ult} is greater than 486.1. However, this result could not be accepted as a realistic estimate of the probability, because it takes no account whatsoever of the uncertainty in estimating the model parameters.

3.4 Raleigh Snowfall Example

The example of Section 3.3 is unusual in providing such a sharp contrast between the maximum likelihood parameter estimation approach and the Bayesian predictive approach, but the general implications of this example are relevant to a variety of problems connected with extreme values. Many extreme value problems arising in applied science are really concerned with estimating probabilities of specific outcomes rather than estimating model parameters, but until recently, this distinction was usually ignored.

As a further example, we consider again the Raleigh snowfall data set of Section 1.1. Specifically, we return to the data set of Table 1 and ask what is the probability, in a single January, of a snowfall equal to or greater than 20.3 inches. Assuming either the Poisson-GPD model (Section 2.2.1) with parameters (λ, σ, ξ), or the equivalent point process approach (Section 2.5) with parameters (μ, ψ, ξ), we can estimate this probability assuming either the maximum likelihood plug-in approach or the Bayesian approach. In either case, it is necessary to choose a specific threshold, confining the estimation to those observations that are above the given threshold.

Fig. 7 shows the predictive probability computed both by the Bayesian formula (3.6) (denoted by B on the figure) or by the maximum likelihood plug-in approach (denoted by M). Both quantities are in turn computed for a variety of different thresholds. For ease of plotting and annotation, the quantity actually plotted is N, where $1/N$ is the predictive probability.

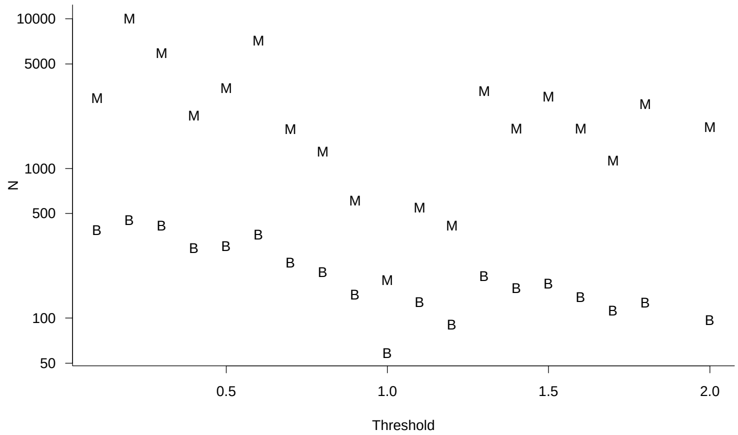


Figure 7. Maximum likelihood (M) and Bayesian (B) estimates of predictive probability $1/N$ for different thresholds

In this case we can see, with both the maximum likelihood and the Bayesian results, that there is a huge dependence on the threshold, but the Bayesian results are all below the corresponding maximum likelihood results (i.e. the Bayesian predictive probability of an extreme event is larger than the maximum likelihood probability) and also arguably more consistent across different thresholds. This does not, of course, prove that the Bayesian estimates perform better than the maximum likelihood estimates, but it does serve to illustrate the contrasts between the two approaches. Most approaches to extreme value analysis keep separate the procedures of estimating unknown parameters and calculating probabilities of extreme events, and therefore by default use the plug-in approach.

The fact that the predictive probabilities, whether Bayesian or maximum likelihood, vary considerably with the somewhat arbitrary choice of a threshold, is still of concern. However, it can be put in some perspective when the variability between predictive probabilities for different thresholds is compared with the inherent uncertainty of those estimates. Confidence intervals for the predictive probability computed

by the delta method (not shown) are typically very wide, while Fig. 8 shows the posterior density of the predictive probability $1/N$ for two of the thresholds, 0.5 and 1, for which the point predictive probabilities are at opposite ends of the spectrum. The substantial overlap between these two posterior densities underlines the inherent variability of the procedure.

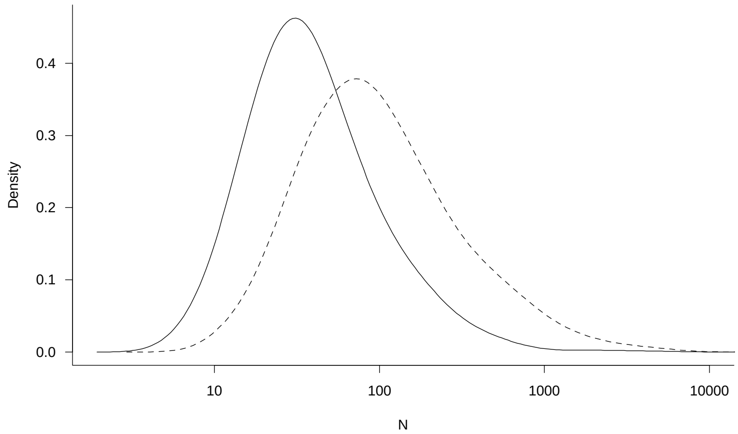


Figure 8. Posterior densities of $1/N$ based on thresholds 1 (solid) and 0.5 (dashed).

In summary, the main messages of this example are:

1. The point estimates (maximum likelihood or Bayes) are quite sensitive to the chosen threshold, and in the absence of a generally agreed criterion for choosing the threshold, this is an admitted difficulty of the approach;
2. The Bayesian estimates of N are nearly always smaller than the maximum likelihood estimates — in other words, Bayesian methods tend to lead to a larger (more conservative) estimate of the probability of an extreme event;
3. The variability among point estimates for different thresholds is less important than the inherent variability of the procedure, based on the standard error in the maximum likelihood case or the posterior density in the Bayesian case.

This example is somewhat unusual in that it represents a very considerable extrapolation beyond the range of the observed data, so it should not be surprising that all the estimates have very high variability. However, the Bayesian results are generally consistent with a return period of between 100 and 200 years, which in turn seems to be consistent with the judgement of most meteorologists based on newspaper reports at the time this event occured.

4. DIAGNOSTICS

The example of Section 3.4 has highlighted one difficulty in applying threshold-based methods: the lack of a clear-cut criterion for choosing the threshold. If the threshold is chosen too high, then there are not enough exceedances over the threshold to obtain good estimates of the extreme value parameters, and consequently, the variances of the estimators are high. Conversely, if the threshold is too low, the GPD may not be a good fit to the excesses over the threshold and consequently there will be a bias in the estimates. There is an extensive literature on the attempt to choose an optimal threshold by, for example, a minimum mean squared error criterion, but it is questionable whether these techniques are preferable in practice to more *ad hoc* criteria, based on the fit of the model to the data. In any case, it is clearly desirable to have some diagnostic procedures to decide how well the models fit the data, and we consider some of these here. The emphasis is on graphical procedures.

4.1 Gumbel Plots

This is the oldest method, appropriate for examining the fit of annual maxima data (or maxima over some other time period) to a Gumbel distribution. Suppose the annual maxima over N years are $Y_1, ..., Y_N$, ordered as $Y_{1:N} \leq ... \leq Y_{N:N}$; then $Y_{i:N}$, for $1 \leq i \leq N$, is plotted against the *reduced value* $x_{i:N}$, where

$$x_{i:N} = -\log(-\log p_{i:N}),$$

$p_{i:N}$ being the ith *plotting position*, usually taken to be $\left(i - \frac{1}{2}\right)/N$.

A straight line indicates good fit to the Gumbel distribution. Curvature upwards or downwards may indicate, respectively, a Fréchet or Weibull distribution. The method is also a useful way to detect outliers.

Examples. Fig. 9(a) is a Gumbel plot based on the annual maxima of the River Nidd river flow series. This is a fairly typical example of a Gumbel plot in practice: although it is not a perfect straight line, there is no systematic evidence of curvature upwards or downwards, nor do there seem to be any outliers. On this basis we conclude that the Gumbel distribution would be a reasonable fit to the data.

Fig. 9(b), taken from Smith (1990), is another example based on annual maximum temperatures (in °C) at Ivigtut, Iceland. Two features stand out: (a) the largest observation seems to be a clear outlier relative to the rest of the data, (b) when this observation is ignored, the rest of the plot shows a clear downward curvature, indicating the Weibull form of extreme value distribution and a finite upper endpoint.

Plots of this nature were very widely used in the early days of the subject when, before automatic methods such as maximum likelihood became established, they were widely used for estimation as well as model checking (Gumbel 1958). This aspect is now not important, but the use of Gumbel plots as a diagnostic device is still useful.

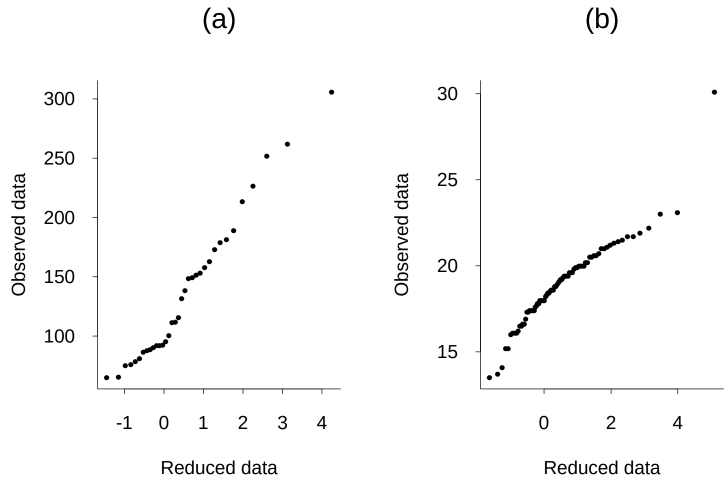


Figure 9. Gumbel plots. (a) Annual maxima for River Nidd river flow series. (b) Annual maximum temperatures in Ivigtut, Iceland.

4.2 QQ Plots

A second type of probability plot is drawn *after* fitting the model. Suppose $Y_1, ..., Y_N$ are i.i.d. observations whose common distribution function is $G(y; \theta)$ depending on parameter vector θ. Suppose θ has been estimated by $\hat{\theta}$, and let $G^{-1}(p; \theta)$ denote the inverse distribution function of G, written as a function of θ. A QQ (quantile-quantile) plot consists of first ordering the observations $Y_{1:N} \leq ... \leq Y_{N:N}$, and then plotting $Y_{i:N}$ against the reduced value

$$x_{i:N} = G^{-1}(p_{i:N}; \hat{\theta}),$$

where $p_{i:N}$ may be again taken as $(i - \frac{1}{2})/N$. If the model is a good fit, the plot should be roughly a straight line of unit slope through the origin.

Fig. 10 illustrates this idea for the Ivigtut data of Fig. 9(b). In Fig. 10(a), the GEV distribution is fitted by maximum likelihood to the whole data set, and a QQ plot is drawn. The shape of the plot — with several points below the straight line at the right-hand end of the plot, except for the final data point which is well above — supports the treatment of the final data point as an outlier. In Fig. 10(b), the same points are plotted (including the final data point), but for the purpose of estimating the parameters, the final observation was omitted. In this case, the plot seems to stick very closely to the straight line, except for the final data point which is an even more obvious outlier than in Fig. 10(a). Taken together, the two plots show that the largest data point is not only an outlier but also an influential data point, i.e. the fitted model is substantially different when the data point is included from when it is not. On the other hand the plot also confirms that if this suspect data point is omitted, the GEV indeed fits the rest well.

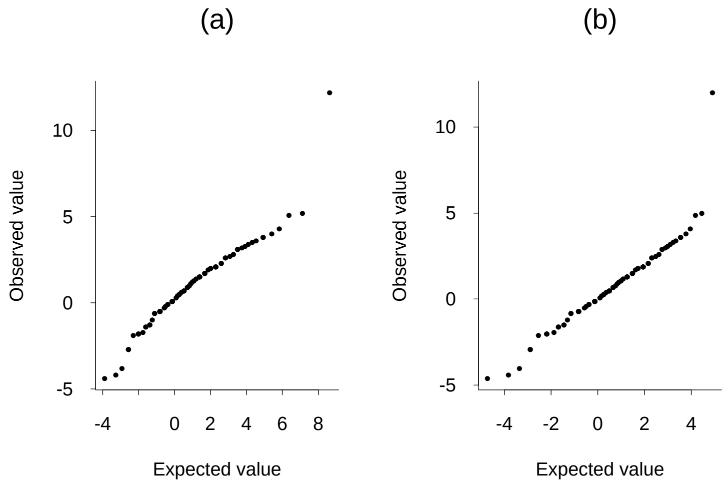


Figure 10. GEV model to Ivigtut data, (a) without adjustment, (b) excluding largest value from model fit but including it in the plot.

Fig. 11, from Davison and Smith (1990), shows QQ plots for the Nidd data introduced in Section 2.5. In this case, $Y_1, ..., Y_N$ are the excess values over a high threshold, to which the GPD is fitted. The entire calculation (model fit followed by QQ plot) is carried out for two thresholds, 70 in plot (a) and 100 in plot (b). Plot (a) shows some strange behaviour, the final two data points below the straight line but before them, a sequence of plotted points above the straight line. No single observation appears as an outlier but the plot suggests that the GPD does not fit the data very well. Plot (b) shows no such problem, suggesting that the GPD is a good fit to the data over threshold 100. Davison and Smith cited this along with several other pieces of evidence to support using a threshold 100 in their analysis. This example serves to illustrate a general procedure, that the suitability of different possible thresholds may be assessed based on their QQ plots, and this can be used as a guide to selecting the threshold.

Fig. 12 shows two examples from insurance data sets. Plot (a) again shows a scatterplot of the oil company data set from Section 1.2, and (b) a QQ plot based on the GPD fitted to all exceedances above threshold 5. The main point here is that although there are reasons for treating the largest two observations as outliers, they are not in fact very far from the straight line — in other words, the data are in fact consistent with the fitted GPD, which in this case is extremely long-tailed (see Section 6 for further discussion). If this interpretation is accepted, there is no reason to treat those observations as outliers. In contrast, plots (c) and (d), taken from another, unpublished, study of oil industry insurance claims, in this case spanning several companies and a worldwide data base, shows the largest observation as an enormous outlier; several alternative analyses were tried, including varying the threshold and performing regression analysis on various covariates (including the size of the spill in barrels, the x coordinate of plot (c)), but none succeeded in "explaining" this outlier. The outlier is the Exxon Valdez oil spill in Alaska, and the largest component of the cost was the punitive damages assessed in litigation (9 billion dollars at the time the analysis was conducted, though the amount was recently reduced to 3 billion dollars on appeal).

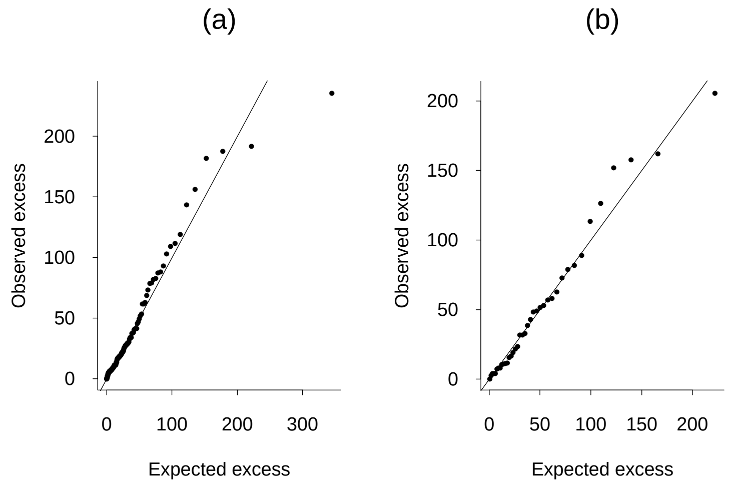


Figure 11. QQ plots for GPD, Nidd data: (a) $u = 70$, (b) $u = 100$.

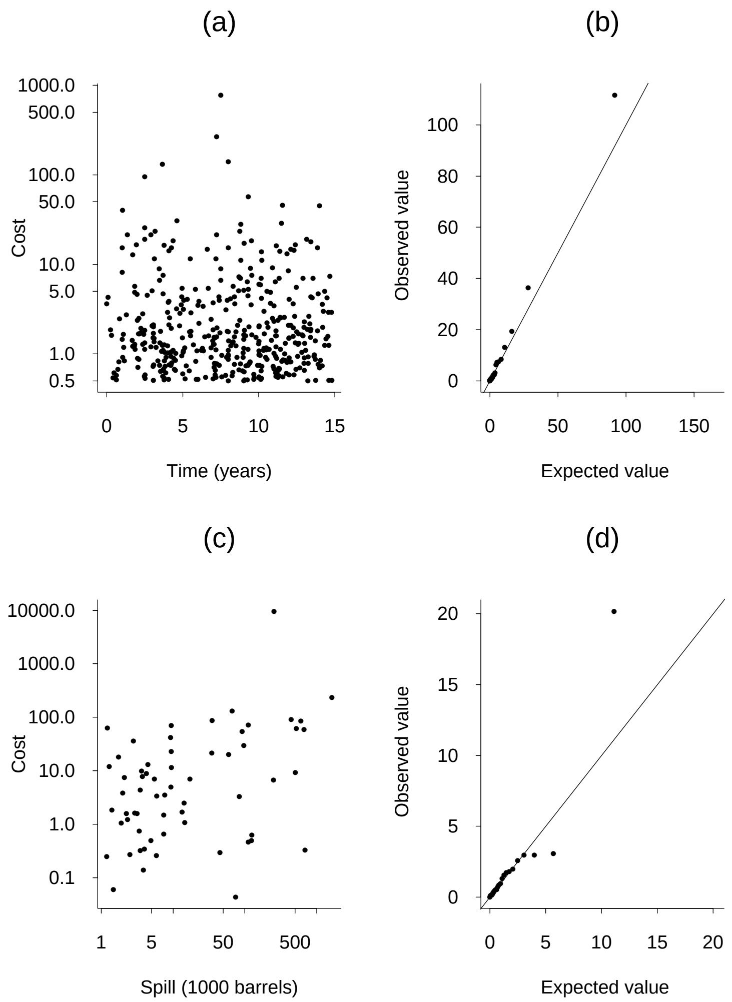


Figure 12. Insurance claims data set: (a) Scatterplot of insurance claims data; (b) QQ plot from GPD fit to insurance claims data; (c) scatterplot of costs of oil spills *vs.* size of spill; (d) QQ plot for oil spills data based on GPD fit with regressors.

The implications of these examples are that observations which first appear to be outliers (as in the first oil company data set) may not in fact be inconsistent with the rest of the data if they come from a long-tailed distribution, whereas in other cases (such as the Ivigtut temperature series and the second oil company example), no amount of fitting different models will make the outlier go away. This is a somewhat different interpretation of outliers from that usually given in statistics; in most statistical applications, the primary interest is in the center of the distribution, not the tails, so the main concern with outliers is to identify them so that they may be eliminated from the study. In an extreme value analysis, it may be important to determine whether the largest observations are simply the anticipated outcome of a long-tailed process, or are truly anomalous. QQ plots are one device to try to make that distinction.

QQ plots can be extended beyond the case of i.i.d. data, for example, to a regression model in which the Y_is are residuals, but it still requires that the residuals have a common distribution. On the other hand, the W plot of Section 4.4 represents a different variant on the idea, in which there is no assumption of homogeneity in the data.

4.3 The Mean Excess Plot

This idea was introduced by Davison and Smith (1990), and is something of an analog of the Gumbel plot for threshold-exceedance data, in the sense that it is a diagnostic plot drawn before fitting any model and can therefore give guidance about what threshold to use.

The mathematical basis for this method is equation (2.9), the key feature of which is that if Y is GPD, then the mean excess over a threshold y, for any $y > 0$, is a linear function of y, with slope $\xi/(1-\xi)$. Thus, we can draw a plot in which the abscissa is the threshold and the ordinate is the sample mean of all excesses over that threshold. The slope of the plot leads to a quick estimate of ξ: in particular, an increasing plot indicates $\xi > 0$, a decreasing plot indicates $\xi < 0$, and one of roughly constant slope indicates that ξ is near 0.

One difficulty with this method is that the sample mean excess plot typically shows very high variability, particularly at high thresholds. This can make it difficult to decide whether an observed departure from linearity is in fact due to failure of the GPD or is just sample variability.

The following Monte Carlo procedure may be used to give a rough confidence band on the plot. Suppose, for some finite u, the true distribution of excesses over u is exactly GPD with parameters (σ, ξ). Suppose the estimates of σ and ξ, based on all exceedances over u, are $\hat{\sigma}$ and $\hat{\xi}$. Also let $\mu(y) = \{\sigma + \xi(y-u)\}/(1-\xi)$ be the theoretical mean excess over threshold for $y > u$, and let $\hat{\mu}(y)$ be the sample mean excess.

A natural test statistic for the GPD assumption is

$$\hat{\mu}(y) - \frac{\hat{\sigma} + \hat{\xi}(y-u)}{1-\hat{\xi}} \tag{4.1}$$

for any given y: this represents the estimated difference between the empirical and theoretical mean excesses at that y.

We can simulate the distribution of (4.1), as follows. For $j = 1, 2, ..., 99$, generate a random sample from the GPD over threshold u, of the same size as the original sample, based on parameters $\hat{\xi}, \hat{\sigma}$. For each j, calculate new MLEs $\hat{\xi}^{(j)}, \hat{\sigma}^{(j)}$ and also the sample mean excess function $\hat{\mu}^{(j)}(y)$. For each u, compute the fifth largest and fifth smallest values of

$$\hat{\mu}^{(j)}(y) - \frac{\hat{\sigma}^{(j)} + \hat{\xi}^{(j)}(y-u)}{1-\hat{\xi}^{(j)}} + \frac{\hat{\sigma} + \hat{\xi}(y-u)}{1-\hat{\xi}} \tag{4.2}$$

as index j ranges over the random samples 1,...,99. Then these values form, for each y, approximate 5% upper and lower confidence bounds on $\hat{\mu}(y)$, if the GPD is correct.

It should be pointed out that this is only a pointwise test, i.e. the claimed 90% confidence level is true for any given y but not simultaneously over all y. Therefore, the test needs some caution in its interpretation

— if the plot remains within the confidence bands over most of its range but strays outside for a small part of its range, that does not necessarily indicate lack of fit of the GPD. Nevertheless, this Monte Carlo procedure can be very useful in gauging how much variability to expect in the mean excess plot.

Fig. 13 shows the mean excess plot, with confidence bands, for the Nidd data, based on all exceedances over thresholds $u = 70$ (plot (a)) and $u = 100$ (plot (b)). The dotted straight line is the estimated theoretical mean excess assuming the GPD at threshold u, and the jagged dashed lines are the estimated confidence bands based on (4.2). Both plots lie nearly everywhere inside the confidence bands, but plot (a) appears to show more systematic departure from a straight line than (b) (note, in particular, the change of slope near $y = 120$), adding to the evidence that threshold 100 is a better bet than 70.

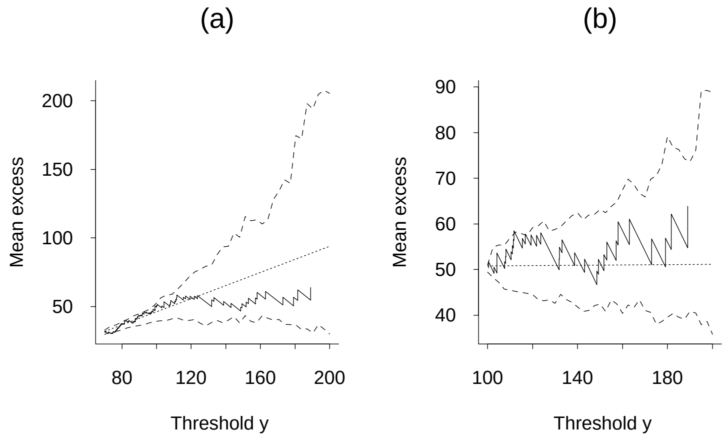


Figure 13. Mean excess over threshold plots for Nidd data, with Monte Carlo confidence bands, relative to threshold 70 (a) and 100 (b).

As another example, we consider three windspeed data sets for cities in North Carolina, based on 22 years of daily windspeed data at Raleigh, Greensboro and Charlotte. For the moment we shall assess these series solely from the point of view of their suitability for a threshold analysis; later (Section 5.2) we shall consider the influence of seasonality and the possibility of long-term trends.

Fig. 14 shows mean excess plots constructed with confidence bands for the three series. The jagged appearance of the plots occurs with all plots of this nature, due to the discretisation of the data. The first plot shows the plot for the Raleigh data computed relative to the threshold $u = 20$ knots. It is clear that there are problems with this plot, since it crosses the upper boundary of the Monte Carlo confidence limits in numerous places. The remaining three plots are all calculated relative to threshold 39.5 (just below 40, to ensure that windspeeds of exactly 40 knots would be included in the analysis) and show no major problems with the fit, though for all three series there are isolated values of the mean excess plot that cross the Monte Carlo confidence limits.

The conclusion in this case is that (for Raleigh at least) 20 knots is clearly too low a threshold for the GPD to be credible model, but for a threshold of 39.5 knots, it seems to be reasonable.

Other examples of mean excess plots are given in Section 5.3.

4.4. Z- and W-statistic Plots

Consider the nonstationary version of the point process model (Section 2.5) with μ_t, ψ_t, ξ_t dependent on t. This is the most general form of threshold model that we have seen so far. We denote the exceedance times $T_1, T_2, ...$ $(T_0 = 0)$ and the corresponding excess values $Y_1, Y_2, ...,$, where Y_k is the excess over the threshold at time T_k.

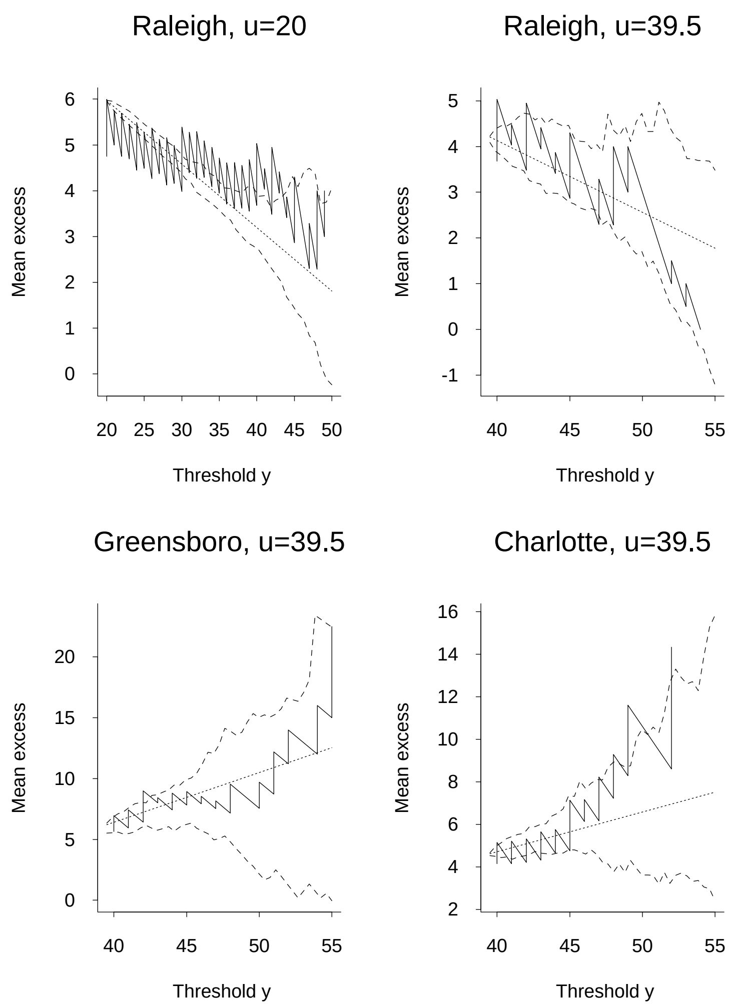


Figure 14. Mean excess plots (with Monte Carlo confidence bands) for windspeed data sets.

The Z statistic is based on intervals between exceedances T_k:

$$Z_k = \int_{T_{k-1}}^{T_k} \lambda_u(t)dt, \quad \lambda_u(t) = \{1 + \xi_t(u - \mu_t)/\psi_t)\}^{-1/\xi_t}. \tag{4.3}$$

The idea is that if $\{T_1, T_2, ..., \}$ are viewed as a one-dimensional point process in time, they form a nonhomogeneous Poisson process with intensity function $\lambda_u(\cdot)$; the transformation then ensures that $Z_1, Z_2, ...,$ are i.i.d. exponential random variables with mean 1. In practice this will only be an approximate result, not exact, because we do not know the true values of $\mu_t,\ \psi_t,\ \xi_t$ and have to use estimates.

The W statistic is based on the excess values:

$$W_k = \frac{1}{\xi_{T_k}} \log\left\{1 + \frac{\xi_{T_k} Y_k}{\psi_{T_k} + \xi_{T_k}(u - \mu_{T_k})}\right\}. \tag{4.4}$$

This is equivalent to a probability integral transformation on the excesses over a threshold: if the model is exact, $W_1, W_2, ...$ are also i.i.d. exponential variables with mean 1. Again, in practice this is only an approximation because the parameters are estimated.

The Z- and W-statistics can be tested in various ways to examine how well they agree with the i.i.d. exponential assumption. As an example, Fig. 15 shows three types of plot computed for the Charlotte

windspeed data (the precise analysis from which these plots were constructed includes seasonal factors and is described in Section 5.2).

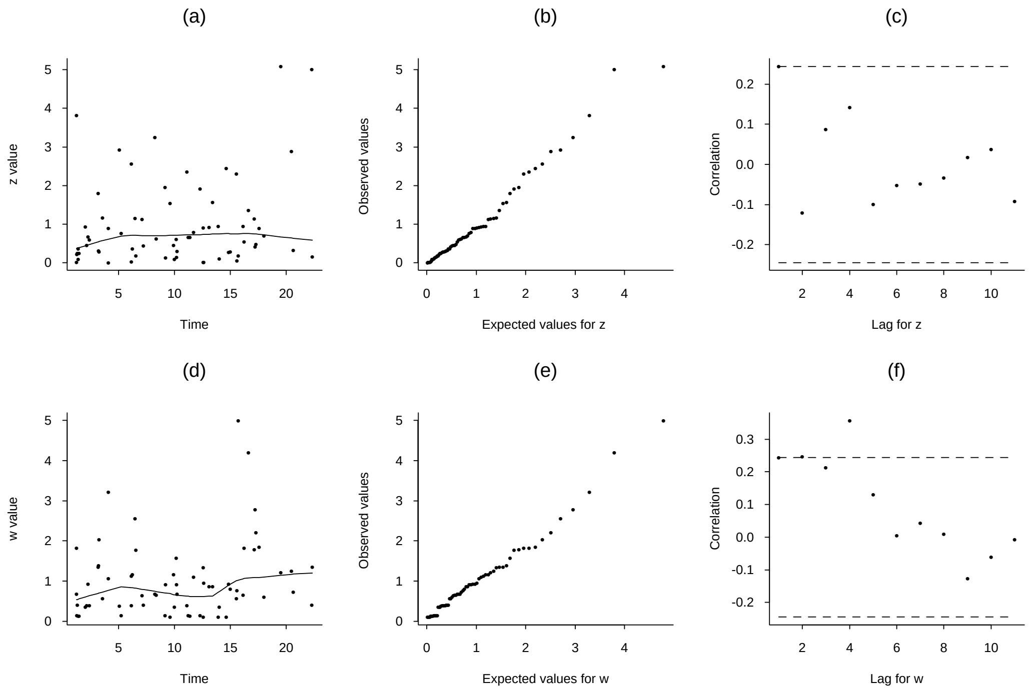


Figure 15. Diagnostic plots based on Z and W statistics for Charlotte seasonal model 1 (Section 5.2). (a,d): plots against time with fitted LOESS curve. (b,e): QQ plots. (c,f): Autocorrelations. The Z-plots are on the top, W-plots on the bottom.

Plots (a) and (d) show the Z and W statistics respectively plotted against time of exceedance, i.e. either Z_k (in plot (a)) or W_k (in plot (d)) is plotted against T_k. The idea here is to observe a possible time trend in the observations. To aid in judging this, a simple fitted curve (using the "lowess" function in S-Plus) is superimposed on the plot. In neither case is there evidence of a systematic trend.

Plots (b) and (e) are QQ plots of the Z and W statistics. Since the hypothesised distribution G is unit exponential, for which $G^{-1}(p) = -\log(1-p)$, this means plotting either $Z_{i:N}$ or $W_{i:N}$ (the ith smallest of N ordered values) against $-\log(1-p_{i:N})$, where $p_{i:N} = (i-\frac{1}{2})/N$. The results in this case show no reason to question the assumption of an exponential distribution with mean 1.

Plots (c) and (f) are plots of the first ten sample autocorrelations for the Z and W values respectively, with approximate confidence bands at $\pm 2/\sqrt{N}$ where N is the sample size. This is a standard plot used in time series analysis and is used here as a indicator of possible serial correlation in either the Z or W series. The only value outside the confidence bands is the fourth autocorrelation coefficient for the W series (plot (f)); in conjunction with moderately high values for the first three autocorrelations, this does perhaps give some suggestion that there is some autocorrelation among the daily excess values of the windspeed series. Given that meteorological phenomena often persist over several days, this is not a surprising conclusion. The threshold methods described so far do not take account of short-term autocorrelations, a theme we return to in Section 8.

The calculation in (4.3) and (4.4) assumes u is a constant. This is another assumption that could be generalised, allowing u to be time-dependent. Thus we could replace u by u_t in (4.3), u_{T_k} in (4.4), to indicate the dependence on time.

5. ENVIRONMENTAL EXTREMES

This section describes three examples of extreme value theory applied to the environment. Section 5.1 is an analysis of ozone extremes, motivated by the problem of determining whether exceedances of an ozone standard are becoming more or less frequent. Section 5.2 treats a similar problem related to windspeed extremes — in this case, the question of exceeding a standard does not arise, but it is a natural question in the context of climate change, to determine whether extreme events are becoming more frequent. Sections 5.3 and 5.4 extend this discussion further, first by asking the same question with respect to rainfall extremes, and then considering how to combine the information from a large number of rainfall stations. Thus, this section uses techniques of spatial statistics as well as extreme value theory.

5.1 Ozone Extremes

As a first example of the use of regression models for extreme value data, we describe an analysis of ozone extremes given originally by Smith and Shively (1995).

The U.S. Environmental Protection Agency (EPA) establishes standards for a number of atmospheric pollutants including ozone. This standard applies to ground-level ozone, which is harmful to human health, and has nothing to do with the stratospheric ozone layer which has the beneficial effect of blocking out certain kinds of radiation. At the time of this study, the EPA standard for ground-level ozone was based on a daily maximum level of 120 ppb (parts per billion). The question of interest here is to try to determine whether the EPA standards have had any beneficial effect — which, for the present study, is interpreted to mean whether there is any downward trend in either the frequency or the size of exceedances over the threshold. However, a major complication in assessing trends is the influence of meteorology. Ozone is produced by various photochemical processes in the atmosphere, and these photochemical processes are highly influenced by the weather (in particular, ozone is likely to be high on windless sunny days in the summer). There is therefore some interest in separating out the trend effects that might be due to genuine trends in the emissions of the so-called ozone precursors — gases such as nitrogen oxides, carbon monoxides, and other pollutants typically produced by industrial processes and by automobiles — and those effects that might be attributed to meteorological variation. The main objective of the present analysis is to build a regression model in which relevant meteorological variables and time are all considered covariates in the regression, so that the coefficient of time may be interpreted as an adjusted time trend after taking the meteorological effects into account.

The data source in this case consists of ozone exceedances from a single station in Houston, 1983–1992. The assumed model has the following components:

(a) The probability of an exceedance of a given level u (in practice taken as 120 ppb, the then-current standard) on day t is $e^{\alpha(t)}$, where

$$\alpha(t) = \alpha_0 + \alpha_1 s(t) + \sum_{j=2}^{p} \alpha_j w_j(t), \tag{5.1}$$

$s(t)$ being the calendar year in which day t falls and $\{w_j(t),\ j = 2, ..., p\}$ the values of $p-1$ weather variables on day t.

(b) Given an exceedance of level u on day t, the probability that it exceeds $u + x$, where $x > 0$, is

$$\{1 + \xi\beta(t)x\}^{-1/\xi}, \tag{5.2}$$

where

$$\beta(t) = \beta_0 + \beta_1 s(t) + \sum_{j=2}^{p} \beta_j w_j(t). \tag{5.3}$$

Note that for (5.1) and (5.3) to make sense, we require $\alpha(t) < 0,\ \beta(t) > 0$, for all t. There is a theoretical possibility that these conditions could be violated, but this was not a practical difficulty in the example under discussion. Also, a further generalisation would also allow ξ to be dependent on covariates — in other words,

replace ξ by $\xi(t)$ where $\xi(t)$ is given by another formula of the form (5.1) or (5.3). This extension was not used in the paper (in fact the authors took $\xi = 0$, as discussed further below).

The meteorological variables were selected after consultation with Texas air control experts who had extensive experience of the kinds of effects that are relevant in the Houston area. A specific list of variables was as follows:

TMAX. Maximum hourly temperature between 6am and 6pm.

TRANGE. Difference between maximum and minimum temperature between 6am and 6pm. This is considered to be a proxy for the amount of sunlight.

WSAVG. Average wind speed from 6am to 6pm. Higher windspeeds lead to lower ozone levels because of more rapid dispersion of ozone precursors.

WSRANGE. Difference between maximum and minimum hourly windspeeds between 6am and 6pm.

NW/NE. Percentage of time between 6am and 6pm that the wind direction was between NW and NE.

NE/ESE. Percentage of time between 6am and 6pm that the wind direction was between NE and ESE.

ESE/SSW. Percentage of time between 6am and 6pm that the wind direction was between ESE and SSW.

SSW/NW. Percentage of time between 6am and 6pm that the wind direction was between SSW and NW.

The wind directions are important for Houston because they determine the level of industrial pollution — for instance, there is a lot of industry to the south of Houston, and ozone levels tend to be higher when the wind direction is in the ESE/SSW sector.

In most analyses, the variable SSW/NW was omitted because of the obvious collinearity with the other wind directions.

The models (5.1)–(5.3) were fitted by numerical maximum likelihood, and standard variable selection procedures were adopted, with variables being dropped from the analysis if their coefficients were not statistically significant. The results are shown in Tables 3 and 4. In (5.2), it was found that the parameter ξ was not significantly different from 0, and of course (5.2) reduces to $e^{-\beta(t)x}$ if $\xi = 0$, so that was the form adopted for the reported analysis.

Variable	Coefficient	Standard Error
$s(t)$	–0.149	0.034
TRANGE	0.072	0.016
WSAVG	–0.926	0.080
WSRANGE	0.223	0.051
NW/NE	–0.850	0.408
NE/ESE	1.432	0.398

Table 3. Coefficient and standard errors for $\alpha(t)$.

Variable	Coefficient	Standard Error
$s(t)$	0.035	0.011
TRANGE	–0.016	0.005
WSAVG	0.102	0.019
NW/NE	0.400	0.018

Table 4. Coefficient and standard errors for $\beta(t)$, assuming $\xi = 0$.

The results show that in both cases the coefficient of $s(t)$ is statistically significant, the sign (negative in Table 3, positive in Table 4) being consistent with the interpretation of an overall decrease in the extreme ozone levels, which was the hoped-for conclusion.

As a comparison, Smith and Shively also fitted the same model for trend alone, ignoring meteorological covariates. In this case, the estimated coefficients of $s(t)$ were –0.069 (standard error 0.030) in $\alpha(t)$ and 0.018 (standard error 0.011) in $\beta(t)$. The coefficients are thus much smaller in magnitude if the model is fitted without any meteorology. This confirms the significance of the meteorological component and shows how the failure to take it into account might obscure the real trend.

Fixing $\xi = 0$ in this analysis would probably not be desirable if the objective was to estimate probabilities of extreme events (or the quantiles associated with specified small return probabilities). This is because the estimates tend to be biased if ξ is assumed to be 0 when it is not, and for this kind of analysis, there is no obvious reason to treat $\xi = 0$ as a null hypothesis value. On the other hand, the main emphasis in this example was on the regression coefficients associated with the extreme events, and especially on the coefficients of $s(t)$ in the expressions for both $\alpha(t)$ and $\beta(t)$, and for that purpose, assuming $\xi = 0$ seems less likely to bias the results. Another possible extension of the analysis would be to allow the threshold u to be time-dependent, though in this example the natural value is 120 ppb because at the time this analysis was originally conducted, that was the U.S. ozone standard.

5.2 Windspeed Extremes

This section illustrates an alternative approach to searching for trends in extreme value data, using the point process approach. The method is applied to the North Carolina windspeed data introduced in Section 4. In this kind of example, it is of interest, especially to insurance companies, to determine whether there is any evidence of a long-term increase in the frequency of extreme weather events. There has been much recent discussion among climatologists (see Section 5.3 for further discussion) that the world's weather is becoming more extreme as a possible side-effect of global warming, so when considering a series of this form, it is natural to look for any possible evidence of a trend. However, another and much more obvious effect is seasonality, and the analysis must also reflect that.

The analysis uses the nonstationary form of the point process model (Section 2.5) in which the GEV parameters are represented by (μ_t, ψ_t, ξ_t) to emphasise the dependence on time t. A typical model is of the form

$$\mu_t = \sum_{j=0}^{q_\mu} \beta_j x_{jt}, \quad \log \psi_t = \sum_{j=0}^{q_\psi} \gamma_j x_{jt}, \quad \xi_t = \sum_{j=0}^{q_\xi} \delta_j x_{jt}, \tag{5.4}$$

in terms of covariates $\{x_{jt},\ j = 0, 1, 2, ...\}$ where we usually assume $x_{0t} \equiv 1$. In the present analysis, we consider models in which only μ_t depends of covariates, so $q_\psi = q_\xi = 0$.

The log likelihood for this model may be derived from the joint density (2.15) and is maximised numerically to obtain estimates of the unknown parameters and their standard errors.

We illustrate the methodology by applying it to the North Carolina windspeed data. The model (5.4) was fitted to the Raleigh data based on exceedances over 39.5 knots. It was restricted to the case $q_\psi = q_\xi = 0$, largely because it simplifies the model not to have too many regression components and it is natural to treat the location parameter μ_t, rather than ψ_t or ξ_t, as the one dependent on covariates.

The covariates x_{jt} were taken to be polynomial functions of time (t, t^2, etc.) and sinusoidal terms of the form $\sin \frac{2\pi k t}{T_0}$, $\cos \frac{2\pi k t}{T_0}$, where $k = 1, 2, ...$ and T_0 represents one year. In practice, no model fitted required sinusoidal terms beyond $k = 2$, and none of the polynomial trend terms were significant. As an example, Tables 5 and 6 give the models fitted to the Raleigh data that involved no covariates (Model 0), and the covariates $\sin \frac{2\pi t}{T_0}$, $\cos \frac{2\pi t}{T_0}$, corresponding to a single sinusoidal curve (Model 1). The difference between negative log likelihoods ($X^2 = 2 \times (114.8 - 103.1) = 23.4$) is clearly significant considering that X^2 has an approximate χ^2_2 distribution when Model 0 is correct. However, when other covariates are added to the model and the likelihood ratio statistics computed, the results are not significant.

Variable	Coefficient	Standard Error
β_0	42.4	0.9
γ_0	1.49	0.16
δ_0	–0.19	0.19

Table 5. Raleigh data, Model 0: no seasonality. NLLH=114.8

Variable	Coefficient	Standard Error
β_0	40.9	0.9
β_1	0.90	1.07
β_2	5.29	1.39
γ_0	1.43	0.13
δ_0	–0.12	0.10

Table 6. Raleigh data, Model 1: Single sinusoidal component. NLLH=103.1

Fig. 16 shows QQ plots for the W statistics for the Raleigh data just considered, and corresponding results for Greensboro and Charlotte. These are based on Model 1 except for plot (d), which is based on Model 0 for Charlotte. Each of plots (a), (b) and (c) shows an excellent fit to the model. Plot (d) is more questionable because the largest two observations appear to be outliers, which is another argument in favor of the seasonal model.

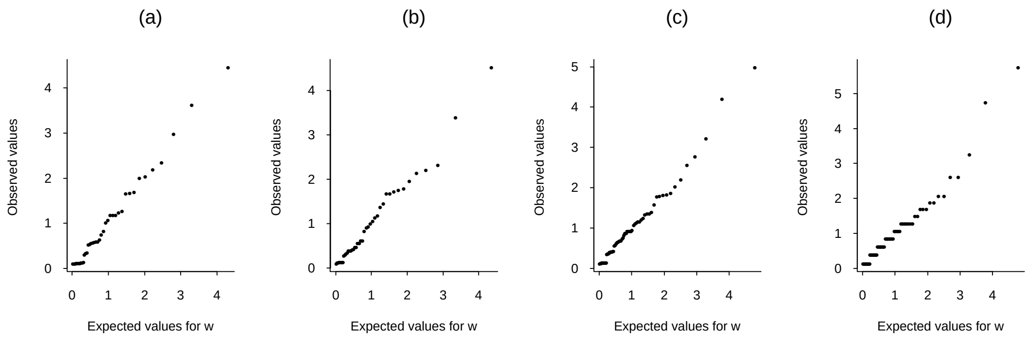


Figure 16. QQ plots of residuals for Raleigh (a), Greensboro (b) and Charlotte (c and d), based on Model 1 except for (d) which uses Model 0.

One conclusion from this analysis is that there do not appear to be overall trends in the series. This is not conclusive in itself because, as we shall see in the context of rainfall extremes in Section 5.4, trends estimated at individual stations tends to vary widely, and it is only when a large number of stations are examined together that an overall positive trend emerges. The other point of debate is whether hurricanes should be treated as separate events from the rest of the data. Meteorologically this makes sense, because a hurricane arises from quite different physical processes from ordinary storm events. Nevertheless, it is not clear how this should affect the data analysis. In the present three series, the two largest windspeeds for each of Greensboro and Charlotte are hurricane events (the largest for Charlotte is Hurriance Hugo, 1989) but there are no hurricanes in this section of the Raleigh data series (Hurricane Fran produced windspeeds up to 79 mph in Raleigh in 1996, but that was after the period covered by the current data set). The results for Greensboro and Charlotte suggest that when a threshold model is fitted to the whole of the data, the

hurricanes do not appear as exceptional outliers. On the other hand, the fitted models for Greensboro and Charlotte have long-tailed distributions (though not given in detail here, both had $\hat{\xi}$ values around 0.2) whereas Raleigh had $\hat{\xi} = -0.19$, indicating a short-tailed distribution. The standard error for $\hat{\xi}$ in Raleigh is also 0.19, indicating that the estimate is not significantly different from 0. Because of their geographical proximity, one would expect the parameters of the three series to be similar, and in Section 5.4 we shall propose a specific model to take that feature into account. For the current data sets, it is as far as we can go to say that the point process model appears to fit adequately to the extremes of all three cities, and that the absence of hurricane events in Raleigh, for the time period of the analysis, does not necessarily mean that the distribution of extreme windspeeds there is significantly different from those of Greensboro and Charlotte.

5.3 Rainfall Extremes

This section is based the preprint of Smith (1999) which examined, from a rather broader perspective than the examples so far, the question of whether there is an overall increasing tendency in extreme rainfall events in the U.S., a popular hypothesis among climatologists.

The data base consisted of 187 stations of daily rainfall data from the Historical Climatological Network (HCN), which is part of the data base maintained by the National Climatic Data Center. Most stations start from 1910 but this analysis is restricted to 1951–1997 during which coverage percentage is fairly constant (Fig. 17).

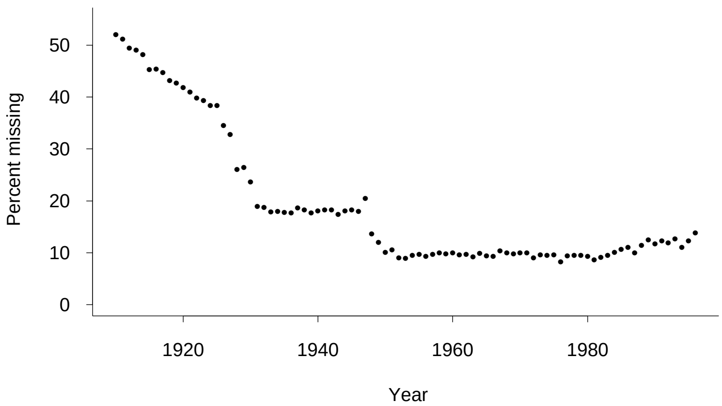


Figure 17. Proportion of missing data over the whole network, for each year from 1910 to 1996.

The strategy adopted in this analysis was to look at four stations in detail, which are widely separated geographically, and then to attempt a spatial analysis over the entire network. The locations of the four stations are shown in Fig. 18.

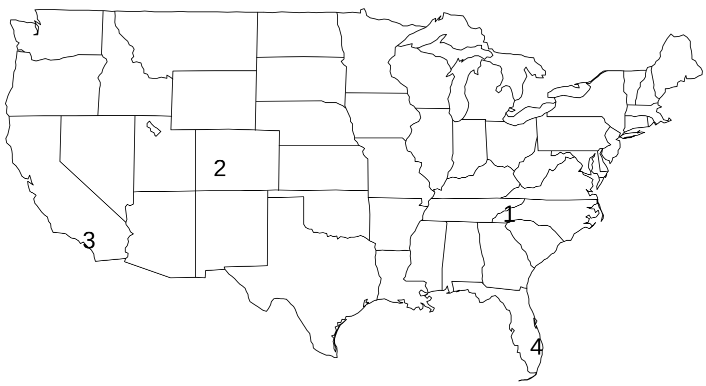


Figure 18. Location of four test stations: 1. Station 319147, Waynesville, NC. 2. Station 53662, Gunnison, CO. 3. Station 49087, Tustin Irvine, CA. 4. Station 80611, Belle Glade, FL.

Figs. 31 and 32 illustrate some rather deceptive features of mean excess plots. In these plots, the units of measurement are $\frac{1}{100}$ inch, and the base threshold relative to which the rest of the plot is calculated has been arbitrarily fixed at the 98th percentile of the raw data at each of the sites. From the simple plots without confidence bands (Fig. 19) one would be tempted to conclude that only Station 2 is reasonably close to a straight line, the others looking very jagged. In fact, when the same plots are drawn with confidence bands (Fig. 20), it becomes clear that Station 2 is the one out of the four for which the excesses over the assumed threshold clearly do *not* fit the Generalised Pareto distribution. Station 2 is Gunnison, Colorado, and the dominant feature of the series is a single very large observation, of around 8 inches, more than three times the second largest observation in the series. When this observation is removed, the mean excess plot looks quite similar to that for the other three stations.

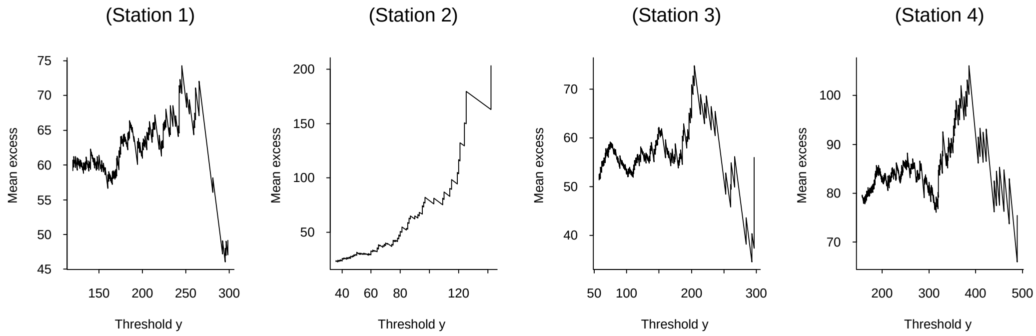


Figure 19. Mean excess plots for four rainfall stations.

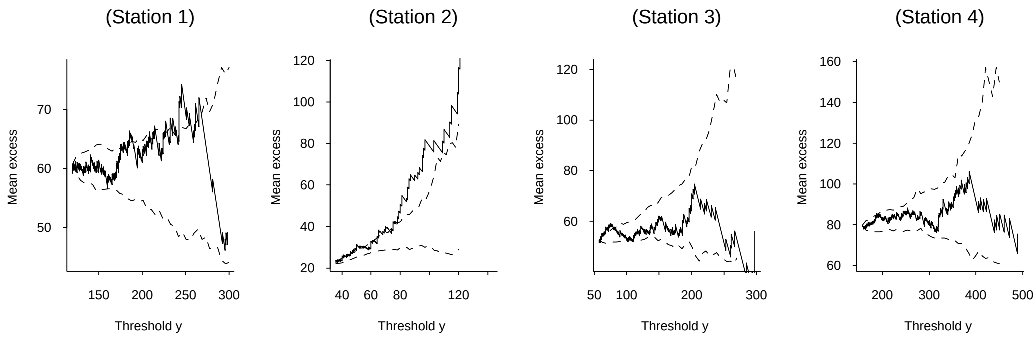


Figure 20. Mean excess plots with confidence bands.

We now consider some possible models to fit to the data at individual stations. All the models are of the nonstationary point process structure of Section 2.5, and we consider two specific model classes:

$$\mu_t = \mu_0 + v_t, \quad \psi_t = \psi_0, \quad \xi_t = \xi_0, \quad \text{(Model 1)}; \tag{5.5}$$

$$\mu_t = \mu_0 e^{v_t}, \quad \psi_t = \psi_0 e^{v_t}, \quad \xi_t = \xi_0, \quad \text{(Model 2)}. \tag{5.6}$$

In either (5.5) or (5.6), v_t is a regression term of the form

$$v_t = \sum x_{tj}\beta_j,$$

where $\{x_{tj}\}$ are known covariates and the $\{\beta_j\}$ are coefficients to be estimated.

For the regression terms x_{tj}, we consider a combination of linear time trends ($x_{tj} = t$), sinusoidal terms of the form $\cos \omega t$, $\sin \omega t$ to represent seasonal effects, and some external signals — as in the analysis of Section 6.4, both the Southern Oscillation Index (SOI) and the North Atlantic Oscillation (NAO) were considered as possible external influences.

Model 2 has some attractiveness because of the following interpretation. Suppose for a moment we ignore the covariates and just use the model to find the n-year return level y_n (i.e. the level which in any one year is exceeded with probability $\frac{1}{n}$) as a function of (μ, ψ, ξ). This is given (approximately) by solving the equation

$$\left(1+\xi\frac{y_n-\mu}{\psi}\right)^{-1/\xi} = \frac{1}{n},$$

leading to the solution

$$y_n = \mu + \psi\left(\frac{n^\xi - 1}{\xi}\right). \tag{5.7}$$

Now consider the case in which μ and ψ are both dependent on time through a function of the form $\mu_t = \mu_0 e^{\beta_1 t}$, $\psi_t = \psi_0 e^{\beta_1 t}$, including the linear covariate on time but ignoring other covariates. Substituting into (5.7), this implies that the n-year return level is itself increasing with time proportional to $e^{\beta_1 t}$. If β_1 is small, this has the interpretation "the extreme rainfall amounts are increasing at a rate $100\beta_1\%$ per year". In contrast, if we take $v_t = \beta_1 t$ in model (5.5), this does not have such a direct interpretation.

For this reason, the analysis uses model 2 as the main model of interest, though it is not clear that this is actually the best-fitting model overall.

Table 7 shows a variety of model fits for Station 1. The covariates tried here included seasonal effects represented by either one or two sinusoidal curves, linear trend LIN, as well as SOI and NAO. The results show: seasonal variation is adequately represented by a single sinusoidal curve; SOI is marginally significant but NAO is not; and LIN is significant. Although our main analysis in this section is for Model 2, the table shows that Model 1 fits the data a little better for this station, raising a possible conflict between choosing the model that has the easiest interpretation and the one that appears to fit the data best.

Model Type	Covariates	NLLH	DF	AIC
2	None	1354.0	3	2714.0
2	Seasonal (1 component)	1350.3	5	2710.6
2	Seasonal (2 components)	1348.8	7	2711.6
2	Seasonal (1 component) + NAO	1349.6	6	2711.2
2	Seasonal (1 component) + SOI	1348.3	6	2708.6
2	Seasonal (1 component) + LIN	1346.9	6	2705.8
2	Seasonal (1 component) + LIN + SOI	1343.5	7	2701.0
1	Seasonal (1 component) + LIN	1344.8	6	2701.6
1	Seasonal (1 component) + LIN + SOI	1341.3	7	2696.6

Table 7. NLLH and AIC values for nine models fitted to Station 1 (Waynesville, North Carolina).

In subsequent discussion we use Model 2 and multiply the parameter β_1 by 100 so that it has the rough interpretation of "percent rise in the most extreme levels per year". For Waynesville, the estimate of this parameter is .074 if SOI is not included, .077 if SOI is included, both with standard error .035, indicating a significant positive trend.

Summary results for the other three stations (each based on Model 2):

In Station 2 (Gunnison, Colorado), there is a large outlier present but it is not "influential" in the sense of affecting the parameter estimates. NAO and SOI are not significant. There is a strong negative linear trend (estimate –.72, standard error .21).

In Station 3 (Tustin Irvine, California) we fit one seasonal component, the SOI is signal stronger than NAO, and the linear trend has estimate .67 without the SOI adjustment, .50 with, each with standard error .26.

In Station 4 (Belle Glade, Florida), there is one seasonal component, and none of the trend terms are significant.

QQ plots based on the Z statistics for the four stations are shown in Fig. 21, and those based on the W statistics are in Fig. 22. For these plots, the outlier in Station 2 has been removed. In general the plots indicate a good fit but there are some features which might justify further investigation, e.g. both the Z and W plots for Station 3 seem somewhat discrepant at the upper end of the plot.

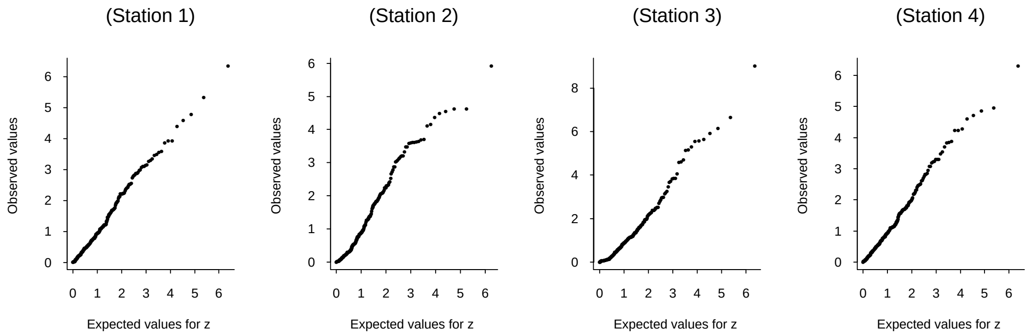


Figure 21. Z-plots for four rainfall stations.

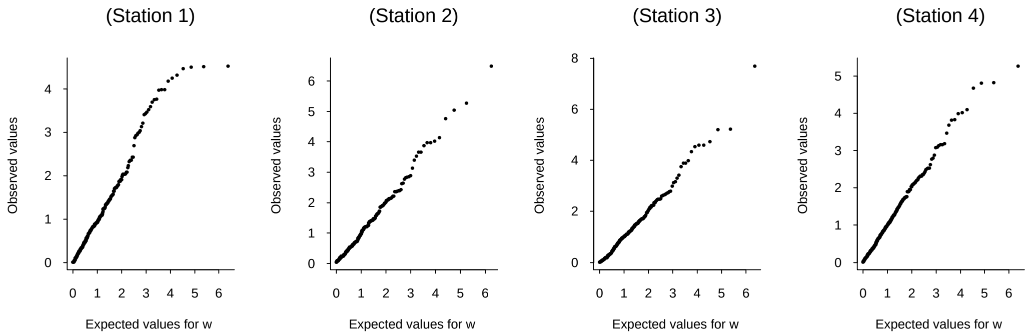


Figure 22. W-plots for four rainfall stations.

5.4 Combining Results Over All Stations

After the preliminary analyses described in Section 5.3, Model 2 was fitted to each of the 187 stations (with successful fits in 184 cases), including a linear trend and two-component seasonal terms, but not SOI or any other external covariate. The threshold for each station was defined as the 98% empirical quantile for the data at that station. The main focus was on the parameter β_1 representing the overall trend, but the analysis also kept track of other parameters including the shape parameter ξ. Apart from the parameter estimates themselves, also computed were the t statistics, calculated by dividing the parameter estimate by its standard error.

	$\hat{\beta}_1$	$\hat{\xi}$
$t > 2$	14 %	40 %
$t > 1$	40 %	73 %
$t > 0$	68 %	88 %
$t < 0$	32 %	12 %
$t < -1$	11 %	2.7 %
$t < -2$	5 %	0.5 %

Table 8. Summary table of t statistics (parameter estimate divided by standard error) for extreme value model applied to 187 stations and 98% threshold. Tabulated values are percentages out of 184 stations analysed.

As an example of the difficulty in determining some overall "significance" of a trend, Table 8 summarises the t statistics for both the β_1 and ξ parameters. For example, if we take $|t| > 2$ as the criterion for a significant result, based on β_1, only 25 of the 184 stations have a significant positive trend, and 10 have a significant negative trend. On the other hand, for 125 stations we have $\hat{\beta}_1 > 0$, compared with only 59 for which $\hat{\beta}_1 < 0$. Thus, it is clear that there is an overall preponderance of stations with positive trend, and more detailed examination of the spatial distribution of $\hat{\beta}_1$ cofficients suggest a fairly random pattern, rather than, for example, the positive coefficients being concentrated in one part of the country and the negative coefficients in another.

The corresponding results for $\hat{\xi}$ in Table 8 show a preponderance of stations for which $\hat{\xi} > 0$. This is also somewhat contradicting conventional wisdom about rainfall distributions, since the gamma distribution is very often used in this context, and the gamma distribution is in the domain of attraction of the Gumbel distribution (Section 2), which would imply $\xi = 0$. The present analysis confirms that most empirical rainfall distributions are somewhat more long-tailed than the gamma.

To integrate the results across the 184 stations, a spatial smoothing technique is adopted, similar to Holland *et al.* (2000).

Suppose there is an underlying "true" field for β_1, denoted $\beta_1(s)$ to indicate the dependence on spatial location s. In other words, we assume (hypothetically) that with infinite data and the ability to determine $\beta_1(s)$ precisely for any location s, we could measure a smooth spatial field. However, in practice we are performing regression analysis and only estimating $\beta_1(s)$, as $\hat{\beta}_1(s)$, at a finite set of locations s. Assuming approximate normality of both the underlying field and the estimates, we can represent this as a two-stage process, with β_1 and $\hat{\beta}_1$ denoting the vector of values at measured locations s,

$$\begin{aligned} \beta_1 &\sim N[X\gamma, \Sigma], \\ \hat{\beta}_1 \ |\beta_1 &\sim N[\beta_1, W]. \end{aligned} \tag{5.8}$$

Combining the two parts of (5.8) into one equation,

$$\hat{\beta}_1 \sim N[X\gamma, \Sigma + W]. \tag{5.9}$$

In these equations, Σ is interpretated as the spatial covariance matrix of the unobserved field β_1 — for the present application, this is taken to be the Matérn covariance function which is popular in spatial analysis (Cressie 1993). Also, $X\gamma$ represents systematic variation in space of the β_1 field: in the present application, after examining various alternatives, the components of the X matrix were taken to be linear and quadratic terms in the latitude and longitude of a station, leading to a quadratic surface as the "mean field". Finally the matrix W is interpreted as the covariance matrix of errors arising from the extreme value regression procedures: this is assumed to be known, and a diagonal matrix, with the variance of $\hat{\beta}_1$ at each station taken as the square of the standard error obtained as part of the extreme value fitting procedure. Equation (5.9) is then fitted to the vector of $\hat{\beta}_1$ coefficients, the regression parameters γ and the unknown parameters of Σ estimated by numerical maximum likelihood. The final step of the fitting procedure uses the fact that once the model parameters are estimated, the relationships (5.8) can be inverted to obtain the conditional distribution of β_1 given $\hat{\beta}_1$: this is applied to obtain a smoothed predictive estimate of the β_1 field, together with estimated prediction covariances, in a manner analogous though not identical to the spatial interpolation procedure known as kriging.

One set of results of this procedure is shown in Fig. 23, which depicts contour and perspective plots of the smoothed surface. Even after smoothing, the surface is somewhat irregular, but it is much smoother than the set of raw $\hat{\beta}_1$ values before any smoothing.

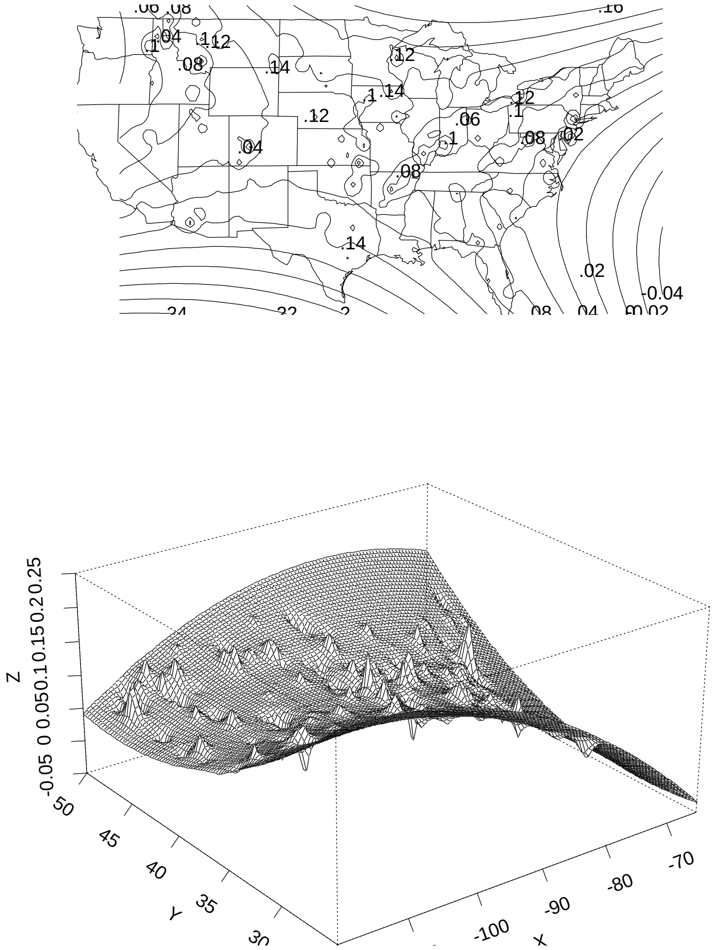


Figure 23. Contour and perspective plots for spatial distribution of rainfall extreme trend coefficients.

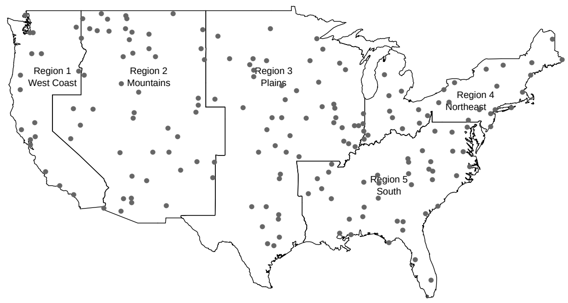


Figure 24. Five regions for regional analysis.

Finally, we consider the computation of regional averages. Fig. 24 (from Grady (2000), who used the same regions in a slightly different context) shows the 187 stations divided into five regions. In Table 9, we give estimated "regional trends", with standard errors. The regional trends are derived by averaging the smoothed β_1 estimates over a very dense grid of points within the region of interest, and the standard errors are the square roots of the estimated variances of those averages, given the $\hat{\beta}_1$ values. These quantities are computed for each of the five regions and also overall, i.e. combining the five regions into one.

Region	Extreme Rainfall Trend	S.E.
1	.055	.024
2	.092	.017
3	.115	.014
4	.097	.016
5	.075	.013
All	.094	.007

Table 9. Regionally averaged trends and standard errors for five regions of USA, and overall.

When interpreted in this way, the results show that each of the five regions has a positive (and statistically significant) trend in the rainfall extremes, but there are also significant differences among the five regions, region 3 in the center of the country having the largest trend and the two coastal regions, region 1 in the west and region 5 in the south-east, having the smallest trends.

6. INSURANCE EXTREMES

This section is concerned with two examples of extremes arising in the insurance industry. Sections 6.1–6.3 extend the discussion of the oil company data set discussed in Section 1. As noted there, after some initial pre-processing of the data, there are 393 claims over a nominal threshold 0.5.

In Section 6.1, we consider the results of the GPD and homogeneous point process model fitted to data above different thresholds. Section 6.2 then extends the discussion to include Bayesian predictive distributions for future losses, drawing on our discussion of Bayesian predictive distributions in Section 3. Section 6.3 is about hierarchical models, in particular arguing the case in favor of treating the different types of claim in Table 2 as coming from separate distributions rather than all arising from a single distribution of claims.

Section 6.4 is on a different theme, where a 75-year series of extreme storms in the U.K. were assessed on the basis of the damage they would have caused under present-day insurance policies. By standardising to present-day values, we take account not merely of price inflation, but of the greatly increased quantity and value of property insured. Such a series is valuable in assessing whether there might have been long-term trends in insurance claims that could be caused by external factors. We again use Bayesian methods to assess risks under various models for a long-term trend.

6.1 Threshold Analyses with Different Thresholds

The first analysis was to fit the GPD to the excesses over various thresholds. The results are shown in Table 10.

A second analysis was to fit the homogenous point process model of Section 2.5, again varying the threshold. In the case of the scale parameter ψ, this is reparametrised as $\log \psi$ for greater numerical stability. The results are in Table 11.

For any given threshold, the models of Tables 10 and 11 are mathematically equivalent, but in comparing results across thresholds, Table 11 is easier to interpret because if the model fits, the (μ, ψ, ξ) parameters will not change with the threshold. In this case, taking into account the standard errors, there seems good agreement among the different parameter estimates across a range of thresholds, confirming the good fit

of the model. This is further confirmed by the diagnostic plots of Fig. 25, which show the Z and W statistics (analogous to Fig. 15) for threshold 5. Based on this, one could recommend threshold 5 as the one to be used for subsequent calculations of extreme event probabilities, though given the near-constancy of parameter estimates across different thresholds, it appears that the results are not very sensitive to the choice of threshold. Note, however, that all the estimated distributions are very long-tailed, with estimates of ξ typically close to 1.

u	N_u	Mean Excess	$\hat{\sigma}$	(S.E.)	$\hat{\xi}$	(S.E.)
0.5	393	7.11	1.02	(0.10)	1.01	(0.10)
2.5	132	17.89	3.47	(0.59)	0.91	(0.17)
5	73	28.9	6.26	(1.44)	0.89	(0.22)
10	42	44.05	10.51	(2.76)	0.84	(0.25)
15	31	53.60	5.68	(2.32)	1.44	(0.45)
20	17	91.21	19.92	(10.42)	1.10	(0.53)
25	13	113.7	33.76	(18.93)	0.93	(0.55)
50	6	37.97	150.8	(106.3)	0.29	(0.57)

Table 10. Fit of the Generalised Pareto distribution to the excesses over various thresholds. The threshold is denoted u, and N_u the number of exceedances over u; we also tabulate the mean of all excesses over u and the maximum likelihood estimates of the GPD parameters σ and ξ. Standard errors are in parentheses.

u	N_u	$\hat{\mu}$	(S.E.)	$\log \hat{\psi}$	(S.E.)	$\hat{\xi}$	(S.E.)
0.5	393	26.5	(4.4)	3.30	(0.24)	1.00	(0.09)
2.5	132	26.3	(5.2)	3.22	(0.31)	0.91	(0.16)
5	73	26.8	(5.5)	3.25	(0.31)	0.89	(0.21)
10	42	27.2	(5.7)	3.22	(0.32)	0.84	(0.25)
15	31	22.3	(3.9)	2.79	(0.46)	1.44	(0.45)
20	17	22.7	(5.7)	3.13	(0.56)	1.10	(0.53)
25	13	20.5	(8.6)	3.39	(0.66)	0.93	(0.56)

Table 11. Fit of the homogenous point process model to exceedances over various thresholds u. Standard errors are in parentheses.

6.2 Predictive Distributions of Future Losses

One question of major interest to the company is how much revenue it needs to put aside to allow for possible losses in some future period. This is a question of predictive inference about future losses, and as noted already in Section 3, there may be good reasons for taking a Bayesian approach to such questions. We approach this here from the point of view of losses over a fixed time period, taken for definiteness to be one year.

Let Y be future total loss in a given year. We write the distribution function of Y as $G(y; \mu, \psi, \xi)$ to indicate the dependence on unknown parameters.

As already noted in Section 3, the traditional frequentist approach to predictive inference is based on the "plug-in" approach in which maximum likelihood estimates are substituted for the unknown parameters,

$$\hat{G}(y) = G(y; \hat{\mu}, \hat{\psi}, \hat{\xi}).$$

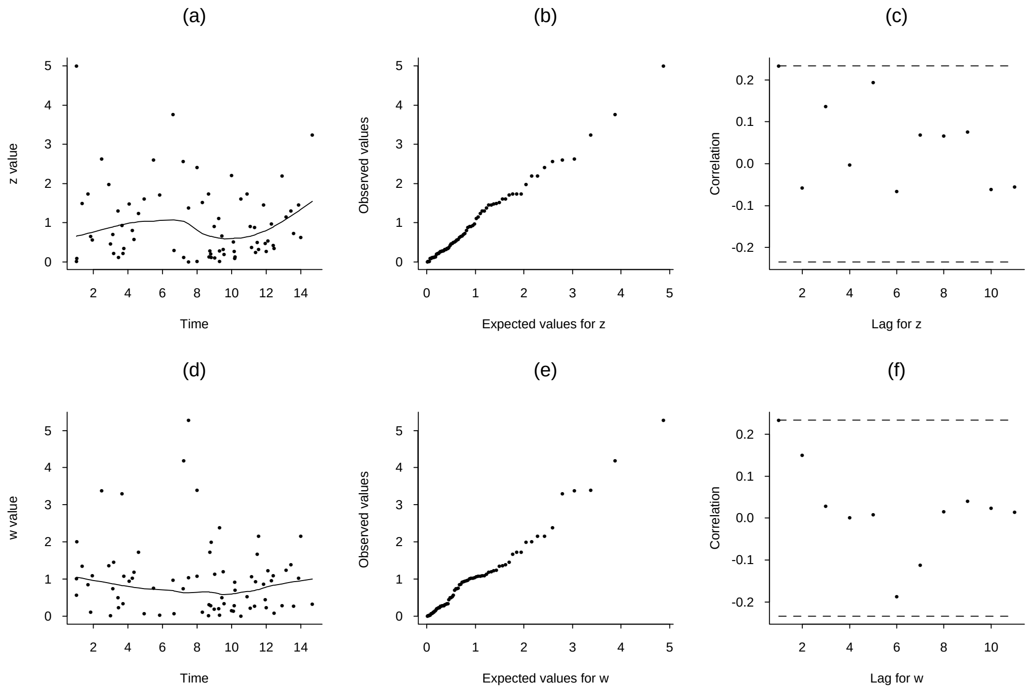


Figure 25. Diagnostic plots based on $u = 5$.

The alternative Bayesian approach leads to

$$\tilde{G}(y) = \int G(y; \mu, \psi, \xi)\pi(\mu, \psi, \xi \mid \mathbf{X})d\mu d\psi d\xi \tag{6.1}$$

where $\pi(\cdot \mid \mathbf{X})$ denotes the posterior density given data $\mathbf{X}$. For the prior density, we take a diffuse but proper prior.

Equation (6.1) is evaluated numerically using a Markov chain Monte Carlo approach. Within this Monte Carlo calculation, the distribution function $G(y; \mu, \psi, \xi)$ is also evaluated by simulation, since it represents the accumulation of many different claims. However, we can assume that the major contribution to the total loss comes from claims over the threshold, so the distribution may be represented as a function of the extreme value parameters μ, ψ and ξ.

In Fig. 26, the posterior densities of μ, ψ and ξ are shown, together with the estimated predictive distribution function for the future losses, calculated from (6.1). To examine the influence of the Monte Carlo error, four independent simulations are performed for each plot. The four estimates within each plot are very close to one another, implying that the Monte Carlo-based error is small.

6.3 Hierarchical Models for Claim Type and Year Effects

In this section, we extend the preceding analysis to allow for the different claim types, and also examine the possibility of a year effect, that might represent some overall trend.

Preliminary analyses of these aspects indicate:

1. When separate GPDs are fitted to each of the 6 main types, there are clear differences among the parameters. (We omit type 7 from this discussion because it contains only two small claims.)
2. The rate of high-threshold crossings does not appear uniform over the different years, but peaks around years 10–12.

To try to take the claim-type effects into account, Smith and Goodman (2000) proposed extending the model into a hierarchical model, of the following structure:

Level I. Parameters $m_\mu,\ m_\psi,\ m_\xi, s_\mu^2,\ s_\psi^2,\ s_\xi^2$ are generated from a prior distribution.

Level II. Conditional on the parameters in Level I, parameters $\mu_1, ..., \mu_J$ (where J is the number of types) are independently drawn from $N(m_\mu, s_\mu^2)$, the normal distribution with mean m_μ, variance s_μ^2. Similarly, $\log \psi_1, ..., \log \psi_J$ are drawn independently from $N(m_\psi, s_\psi^2)$, $\xi_1, ..., \xi_J$ are drawn independently from $N(m_\xi, s_\xi^2)$.

Level III. Conditional on Level II, for each $j \in \{1, ..., J\}$, the point process of exceedances of type j is a realisation from the homogeneous point process model with parameters $\mu_j,\ \psi_j,\ \xi_j$.

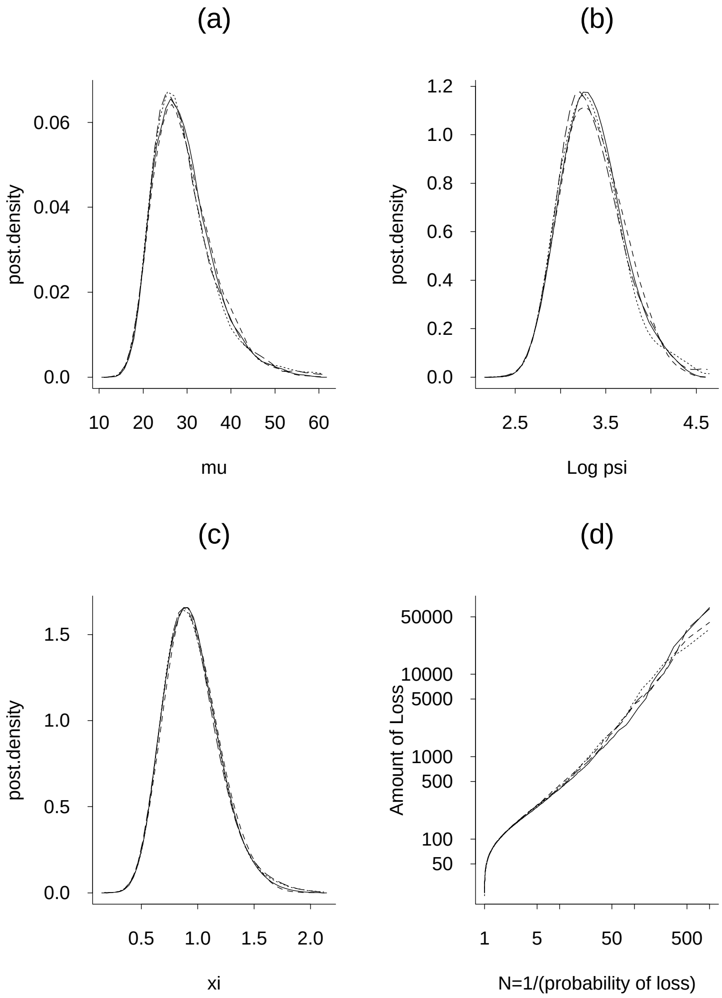


Figure 26. Estimated posterior densities for the three parameters, and for the predictive distribution function (6.1). Four independent Monte Carlo runs are shown for each plot.

This model may be further extended to include a year effect. Suppose the extreme value parameters for type j in year k are not μ_j, ψ_j, ξ_j but $\mu_j + \delta_k, \psi_j, \xi_j$. In other words, we allow for a time trend in the μ_j parameter, but not in ψ_j and ξ_j. We fix $\delta_1 = 0$ to ensure identifiability, and let $\{\delta_k,\ k > 1\}$ follow an AR(1) process:

$$\delta_k = \rho \delta_{k-1} + \eta_k, \quad \eta_k \sim N(0, s_\eta^2) \tag{6.2}$$

with a vague prior on (ρ, s_η^2).

These models are estimated by Markov chain Monte Carlo methods, an extension of the Monte Carlo methods used for computing posterior and predictive densities for single distributions. Fig. 27 shows boxplots based on the Monte Carlo output, for each of $\mu_j, \log \psi_j, \xi_j,\ j = 1, ..., 6$ and for $\delta_k,\ k = 2, ..., 15$. Taking the

posterior interquartile ranges (represented by vertical bars) into account, it appears that there are substantial differences among the six claim types for the μ parameter, and to a lesser extent for the ψ parameter. In contrast, the six types seem homogeneous in the ξ parameter, though it is notable that the posterior means of $\xi_1, ..., \xi_6$ are all in the range 0.7–0.75, compared with a value close to 1 when the six types are combined into one distribution, as in Fig. 26. This shows one effect of the disaggregation into different types — when the types of claims are separated, the apparent distributions are less long-tailed than if the types are aggregated. The effect this has on the predictive distributions will be seen later (Fig. 29). In contrast, the estimates of the δ_k parameters seem fairly homogeneous when the posterior interquartiles ranges are taken into account, suggesting that trends over the years are not a significant feature of this analysis.

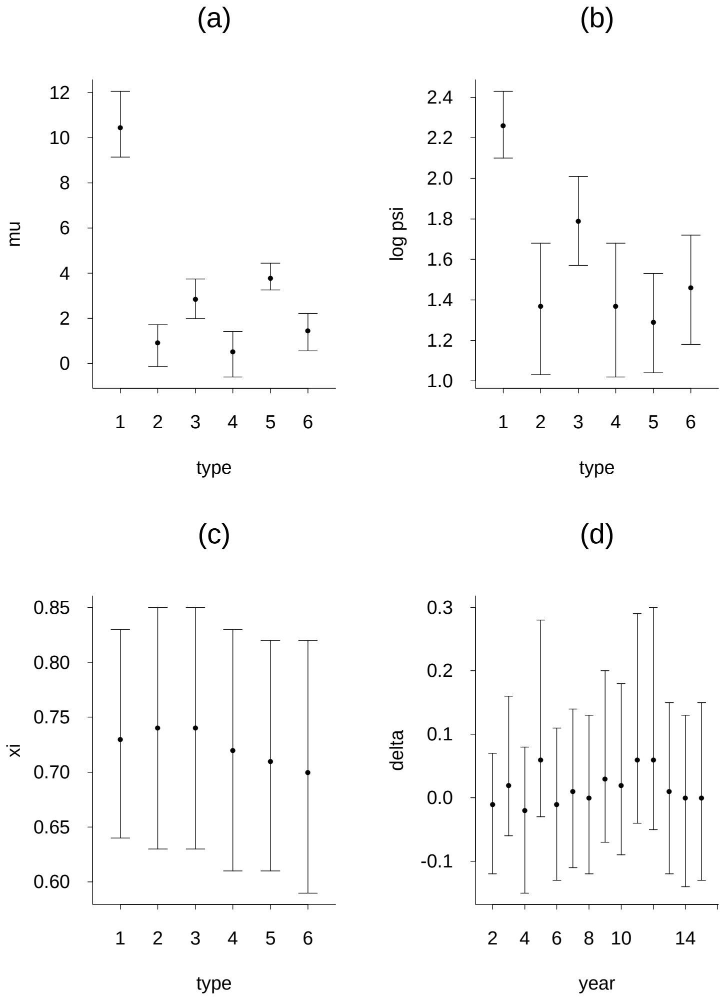


Figure 27. Boxplots for the posterior distributions of μ_j, $\log \psi_j$, ξ_j $(j = 1, ..., 6)$ and for δ_k $(k = 2, ..., 15)$, based on the hierarchical model. The central dots represent posterior means and the horizontal bars the first and third quartiles of the posterior distribution, taken from the Monte Carlo output.

Fig. 28 shows the estimated posterior density of ρ from (6.2), which is fairly flat over the region of stationarity $(-1, 1)$. Finally, Fig. 29 shows the estimated loss curves for future total loss over a one-year period, similar to (6.1) but integrating over all the unknown parameters in the hierarchical model. Four loss curves are computed: curve A based on the homogenous model with all data combined; curve B taking into account the claim-type effects but not year effects; curve C taking into account both claim-type and year effects; and curve D which is computed the same way as curve C, but omitting the two outliers which were mentioned in the earlier discussion. The sharp contrast between curves A and B highlights the advantages of including claim-type effects: we obtain much less extreme predicted losses, a consequence of the fact

that the data appear less long-tailed if the claim types are separated than if they are combined. Including the year effects (curve C) makes little difference to the predicted curve for a typical future year. Curve D was included largely because there was some interest within the company in separating out the "total loss" events (recall from the earlier discussion that the two largest claims in this series were the only claims that represent the total loss of a facility). There is a substantial difference between curves C and D, highlighting that these very extreme events do have a significant influence on future predictions — from the company's point of view, a predicted curve that includes total losses can be expected to differ from one that excludes total losses, and this is reflected in the figure here.

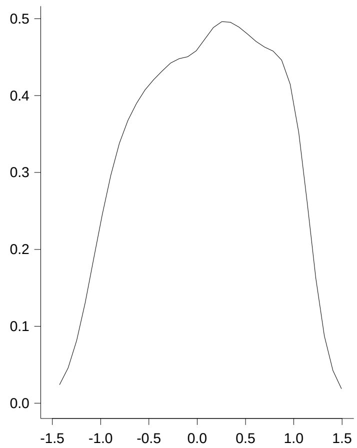


Figure 28. Posterior density of ρ.

6.4 Analysis of a Long-term Series of U.K. Storm Losses

This example is based on a preliminary analysis of a data set constructed by the U.K. insurance company Benfield-Greig. The objective of this analysis is to study long-term trends in insurance claims. However, it is recognised that there are many reasons, besides possible climate-change effects, why insurance claims are rising over time — inflation is an obvious reason, but also, with more property being insured, and insured to higher values in real terms than in earlier times, it is inevitable that insurance claims will increase with time as well. The interest here, however, lies in separating out the effects that might possibly be due to long-term climate change. With this in mind, Benfield-Greig compiled a list of 57 historical storm events in the period 1920–1995, and assessed their insurance impact based on the likely total loss due to each storm that would have occured if the same storm event had occurred in 1998. If *this* series contains a trend, it seems reasonable to conclude that it is due to climatic influences rather than either inflation or changes in insurance coverage.

The analysis adopted here is again of the point process form (5.4) with $q_\psi = q_\xi = 0$, and with μ_t modelled as a function of time t through various covariates:

- Seasonal effects (dominant annual cycle);
- Polynomial terms in t;
- Nonparametric trends;
- Dependence on climatic indicators such as SOI, NAO (discussed later).

Other models in which ψ_t and ξ_t depend on t were also tried but did not produce significantly different results.

Fig. 30 shows scatterplots of the losses against (a) year and (b) day within year — the latter plot shows an obvious and expected seasonal effect but the former is somewhat ambiguous as to whether there is a trend. Fig. 31 shows a mean excess plot, with Monte Carlo confidence bands, to the raw data assuming no trend — since the slope of the fitted straight line is very near 0, this appears fully consistent with an

exponential distribution for exceedances over a high threshold (as noted in Section 4.3, the slope of the mean excess plot may be used as a diagnostic for the sign of ξ).

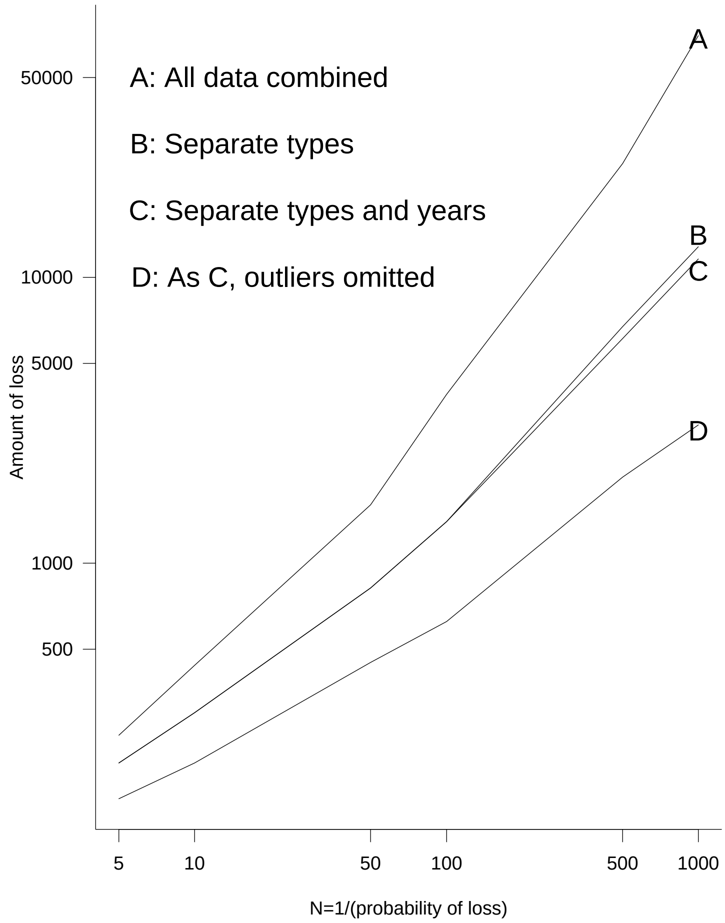


Figure 29. Computations of posterior predictive distribution functions (plotted on a log-log scale) corresponding to the homogenous model (curve A) and three different versions of the hierarchical model.

To fit the data, a number of different trend effects were included. Seasonal effects were modelled by a simple sinusoid of period one year (higher-order harmonics were tested but found not to be significant). Polynomial trends of various orders were tried, the best fit being a cubic curve, and also various nonparametric representations of the long-term trend, using spline functions with various degrees of freedom. Also tried were some meteorologically-based covariates that are believed to influence British storm events: the Southern Oscillation Index (pressure difference between Darwin and Tahiti, henceforth abbreviated to SOI) is very widely used as a measure of the strength of the El Niño effect, but more relevant to British meteorology is the North Atlantic Oscillation or NAO (pressure difference between Iceland and the Azores). A series of different models, with the corresponding values of the negative log likelihood (NLLH) and Akaike Information Criterion (AIC) are given in Table 12. This confirms that the seasonal effect and the NAO are both statistically significant, though not the SOI. The spline trend used here has 5 degrees of freedom and fits the data slightly worse than the cubic trend which has only 3 degrees of freedom; however, given the common-sense conclusion that a nonparametric representation of the long-term trend is likely to prove more satisfactory than the probably accidental fit of a polynomial function, we retain both trend terms for future study.

For the model including the seasonal effect and NAO only, Fig. 32 shows the diagnostic plots based on the Z and W statistics, analogous to our earlier plots in Figs. 15 and 25. There is no evidence in these plots against the fit of the proposed model.

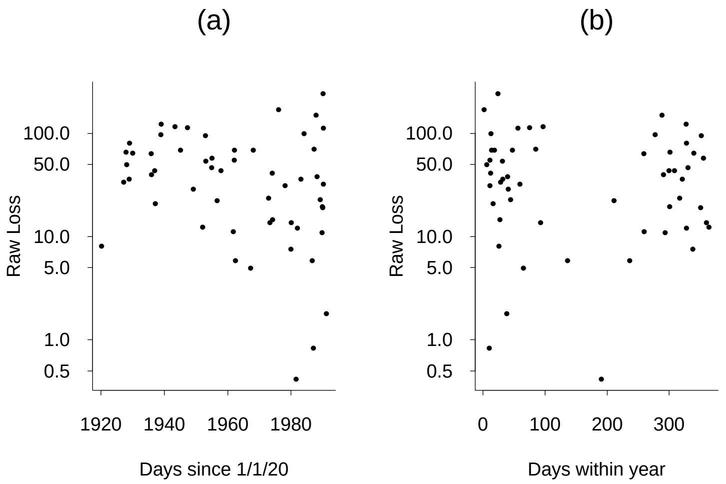


Figure 30. Plots of estimated storm losses against (a) time measured in years, (b) day within the year.

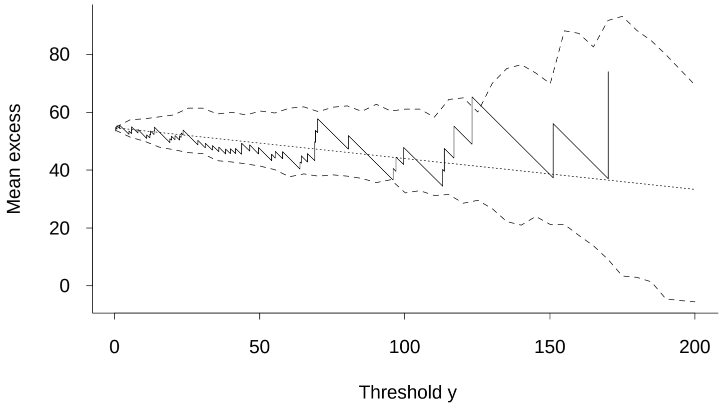


Figure 31. Mean excess plot with Monte Carlo confidence bands.

Model	$p =$ number of parameters	NLLH	AIC
Simple GPD	3	312.5	631.0
Seasonal	5	294.8	599.6
Seasonal + cubic	8	289.3	594.6
Seasonal + spline	10	289.6	599.2
Seasonal + SOI	6	294.5	601.0
Seasonal + NAO	6	289.0	590.0
Seasonal + NAO + cubic	9	284.4	586.8
Seasonal + NAO + spline	11	284.6	591.2

Table 12. NLLH and AIC values (AIC=2 NLLH + 2 p) for several models.

To assess the impact of the trend terms on return levels, Fig. 33(a) shows the 10-year, 100-year and 1000-year return levels for January, under a model with both seasonal and NAO effects. To avoid trying to represent both the seasonal and NAO effects on a single graph, the plot here just represents the loss curve for January. The 10-year return level is here defined as the value that is exceeded on any one day with probability 1/3652, and similar definitions for 100-year and 1000-year return levels. The plot brings out the differences among the 10-year, 100-year and 1000-year return levels, but taking just the middle one of these and plotting it along with its confidence limits in Fig. 33(b) makes clear that the uncertainty of these estimates is higher than the differences among the three curves. For models in which either a cubic or a spline-based trend was included, the corresponding curves are shown as Figs. 34 and 35. Both curves show a dip in the loss curve in the middle of the century, followed by a rise in the period following 1975. Fig. 36 is similar to Fig. 33 but computed for July — of course, the return levels are much lower in this case. Taken together, these plots illustrate the influence of different types of trend, demonstrating that NAO has a strong influence on the return levels but also that there seem to be persistent long-term trends.

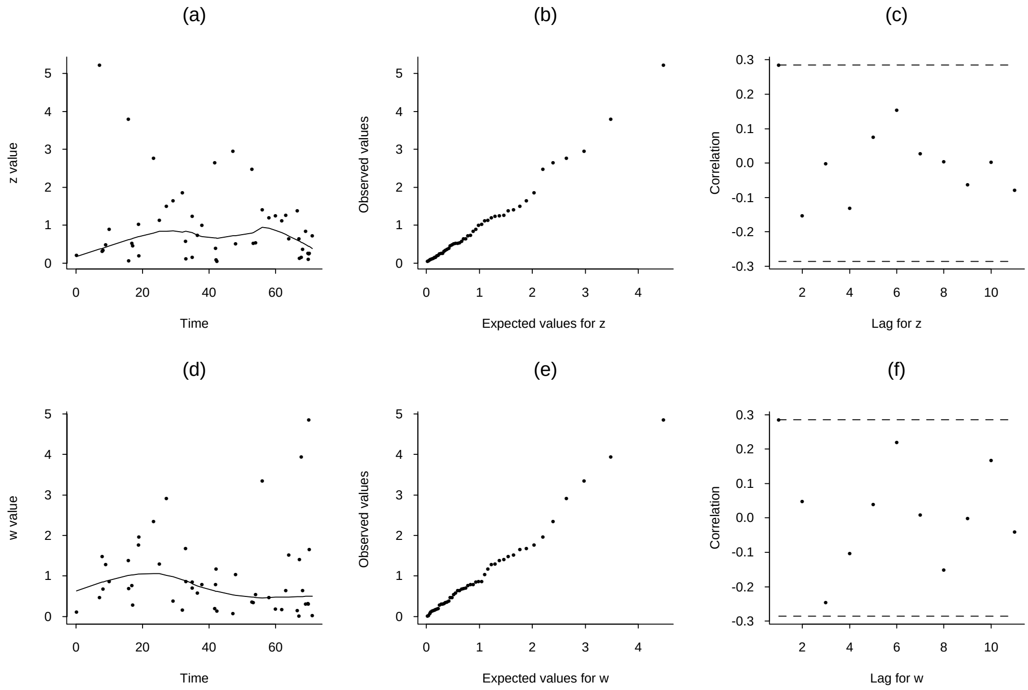


Figure 32. Diagnostic plots based on the Z (plots a,b,c) and W (plots d,e,f) statistics.

Finally, we consider the question that is of most interest to the insurance industry — the likelihood of possibly greater losses in the future. The largest value in the present data set was the January 1990 windstorm which caused damage across a very large area of the U.K. (worse than the storm of October 1987, which involved stronger winds, but was confined to a much smaller geographic area). We consider this question by computing the return period associated with the 1990 event, calculated under various models (Table 13). The first row of this table assumes the simple Pareto distribution — this distribution does not fit the data at all, since as already seen from Fig. 31, the true distribution of the tail is much closer to exponential than Pareto. The return period under the exponential distribution, fitted without any seasonality or trend, is in the second row of the table. Fitting the GPD instead of the Pareto, and then adding a seasonal effect (but for the return level calculations, combining the entire year into a single calculation) has the effect of reducing the estimated return period, compared with the exponential model, but not to a very great extent. The next three rows take the NAO effect into account, first for a "high NAO" year, then for a "low NAO" year, and finally for "random NAO" — in other words, NAO is accounted for in the model but for the return period calculations, it is averaged over the distribution of NAO. In all cases, the quantity actually being estimated is the probability of a loss over the 1990 value occurring in one year; the cited "return period"

is the reciprocal of this. The final row of this table represents the collective wisdom of a group of British insurance industry executives at a workshop in Cambridge, who were shown this table and then asked to give their own estimate of the likelihood of a recurrence of the 1990 event. Whether they were being more pessimistic or simply more realistic than the statistical analysis is a matter on which the reader is left to decide his or her own conclusion.

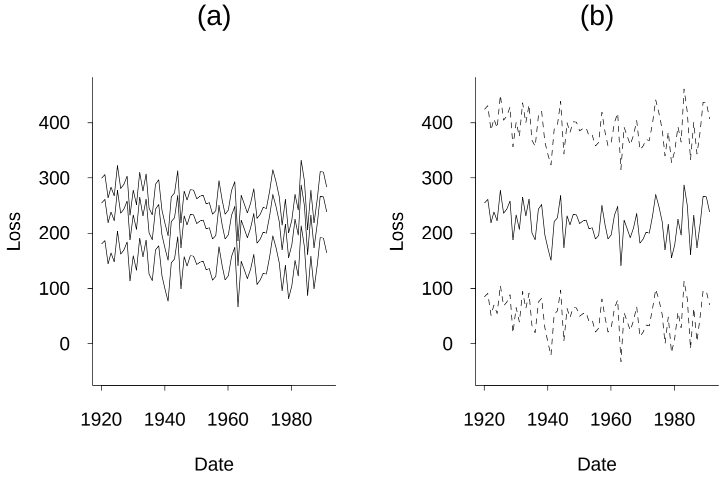


Figure 33. (a) Estimates of 10-year (bottom), 100-year (middle) and 1000-year (top) return levels based on the fitted model for January, assuming long-term trend based on NAO. (b) 100-year return level with confidence limits.

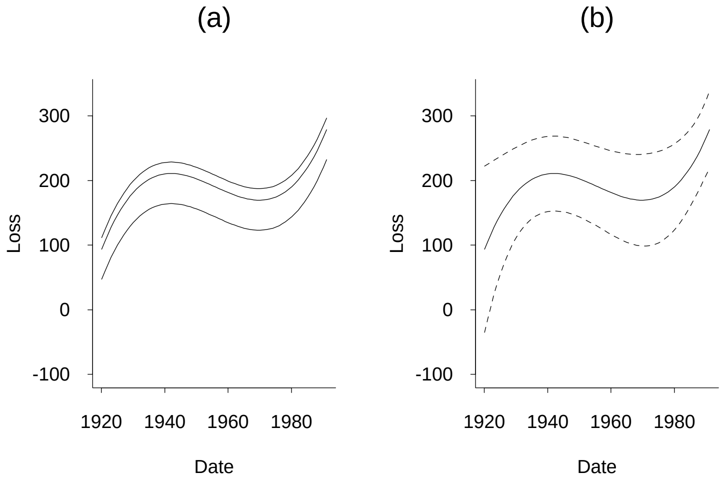


Figure 34. (a) Estimates of 10-year (bottom), 100-year (middle) and 1000-year (top) return levels based on the fitted model for January, assuming long-term trend based on a cubic polynomial. (b) 100-year return level with confidence limits.

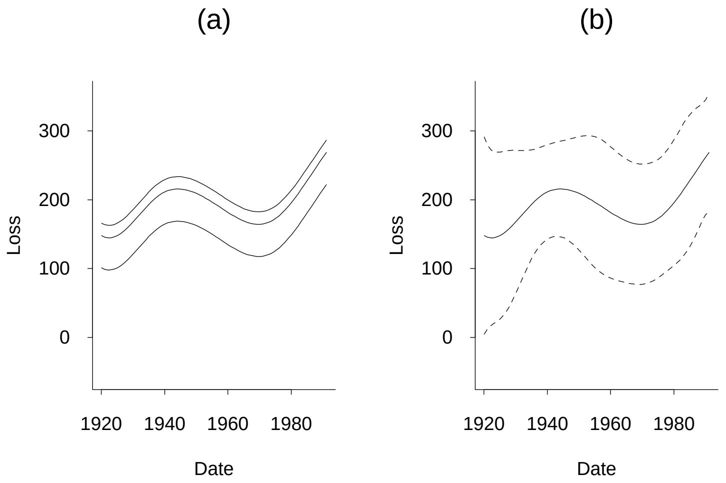


Figure 35. (a) Estimates of 10-year (bottom), 100-year (middle) and 1000-year (top) return levels based on the fitted model for January, assuming long-term trend based on a cubic spline with 5 knots. (b) 100-year return level with confidence limits.

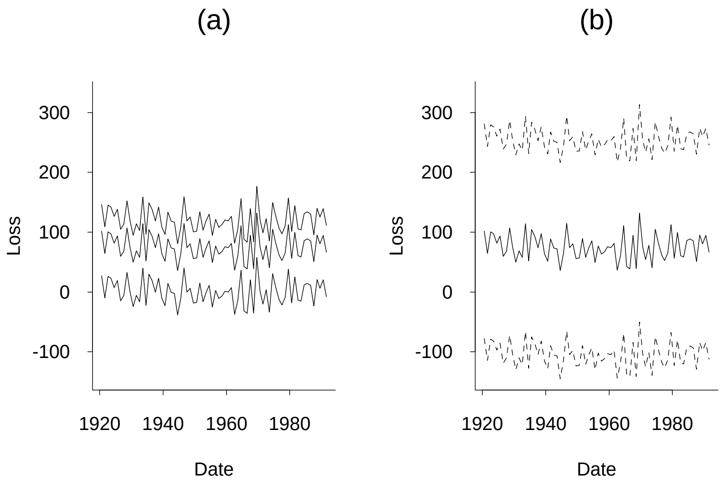


Figure 36. (a) Estimates of 10-year (bottom), 100-year (middle) and 1000-year (top) return levels based on the fitted model for July, assuming long-term trend based on NAO. (b) 100-year return level with confidence limits.

Model	Return Period (years)
Pareto	18.7
Exponential	487
Generalised Pareto	333
GPD with seasons	348
GPD, High NAO year	187
GPD, Low NAO year	1106
GPD with random NAO	432
Current industry estimate?	<50?

Table 13. Estimated return periods (computed from maximum likelihood estimates) associated with the 1990 storm event for several different models.

7. MULTIVARIATE EXTREMES AND MAX-STABLE PROCESSES

Multivariate extreme value theory is concerned with the joint distribution of extremes in two or more random variables. Max-stable processes arise from the extension of multivariate extreme value theory to infinite dimensions. Although there is by now a large literature on multivariate extreme value theory (see, e.g. the chapter by Fougères (2002) in the present volume), there has been much less work on statistical approaches to max-stable processes. In the present section we review these ideas and suggest a particular approach based on so-called M4 processes. In Section 8, we consider a possible application of this approach to financial time series.

7.1 Multivariate Extremes

Suppose $\{X_i = (X_{i1}, ..., X_{iD}),\ i = 1, 2, ...,\}$ are i.i.d. multivariate random vectors, and define vector maxima $M_n = (M_{n1}, ..., M_{nD})$ where $M_{nd} = \max\{X_{id} :\ 1 \le i \le n\}$ for $1 \le d \le D$. Suppose there exist normalising constants $a_{nd} > 0,\ b_{nd}$ such that

$$\frac{M_{nd} - b_{nd}}{a_{nd}}$$

converges in distribution to a nondegenerate limit for each d. In its original form, multivariate extreme value theory was concerned with limiting forms of the joint distribution, i.e. with results of the form

$$\Pr\left\{\frac{M_{nd} - b_{nd}}{a_{nd}} \le y_d,\ 1 \le d \le D\right\} = H(y_1, ..., y_D), \tag{7.1}$$

where H is some nondegenerate D-dimensional distribution function. Of course, it is a consequence of the theory of Section 2 that the marginals of H must be one of the three types of univariate extreme value distributions, but our interest here is in the kinds of joint distributions that can arise.

Given that we have already developed an extensive theory for fitting univariate distributions above a high threshold, we can use this to transform the univariate marginal distributions any way we want. The results we are going to develop are most conveniently expressed if we assume the marginal distributions are unit Fréchet, i.e.

$$\Pr\{X_{id} \le x\} = e^{-1/x} \text{ for each } d, \tag{7.2}$$

so henceforth we assume (7.2). Note also that in this case, we may take the normalising constants to be $a_{nd} = n,\ b_{nd} = 0$.

In practice, we may not be able to achieve (7.2) exactly, but we can perform a probability integral transformation based on the distribution fitted above a high threshold. Suppose we select a threshold u_d for the dth component, and fit the GPD to the excesses of X_{id} above u_d. Then we can estimate parameters $\lambda_d,\ \sigma_d,\ \xi_d$ such that

$$\Pr\{X_{id} \ge u_d\} = \lambda_d, \quad \Pr\{X_{id} \ge u_d + y \mid X_{id} \ge u_d\} = \left(1 + \xi_d \frac{y}{\sigma_d}\right)^{-1/\xi_d},$$

valid over the range where $1 + \xi_d y/\sigma_d > 0$. In that case, the transformation

$$Z_{id} = \left[-\log\left\{1 - \lambda_d\left(1 + \xi_d \frac{X_{id} - u_d}{\sigma_d}\right)^{-1/\xi_d}\right\}\right]^{-1} \tag{7.3}$$

achieves the result

$$\Pr\{Z_{id} \le z\} = e^{-1/z}, \quad z \ge \{-\log(1 - \lambda_d)\}^{-1}. \tag{7.4}$$

With the transformation to unit Fréchet margins, the key property that characterises multivariate extreme value distributions is *max-stability.* Suppose, for each i, $(Y_{i1}, ..., Y_{iD})$ is an independent random vector with distribution function H. For each d, we know that

$$Y^*_{nd} = \frac{1}{n} \max\{Y_{1d}, ..., Y_{nd}\}$$

has the same univariate distribution as Y_{1d}. The property that characterises max-stability is that the same thing should be true for the multivariate distributions:

$$\Pr\{Y^*_{n1} \le y_1, ..., Y^*_{nD} \le y_D\} = H(y_1, ..., y_D). \tag{7.5}$$

Equation (7.5) is equivalent to

$$H(ny_1, ..., ny_D)^n = H(y_1, ..., y_D) \tag{7.6}$$

so (7.6) may be taken as the definition of a multivariate extreme value distribution with unit Fréchet margins. There is an extensive literature on the characterisations and properties of multivariate extreme value distributions, which has been very well reviewed by Fougères (2002).

7.2 Max-stable Processes

Consider now the case of a discrete-time stochastic process, i.e. a dependent sequence $\{Y_i,\ i = 1, 2, ...\}$. If we are interested in the extremal properties of this process, then for the same reasons as with multivariate extremes, it suffices to consider the case where all the marginal distributions have been transformed to unit Fréchet. Such a process is called *max-stable* if all the finite-dimensional distributions are max-stable, i.e. for any $n \ge 1,\ r \ge 1$

$$\Pr\{Y_1 \le ny_1, ..., Y_r \le ny_r\}^n = \Pr\{Y_1 \le y_1, ..., Y_r \le y_r\}.$$

In fact we are interested in D-dimensional processes so it makes sense to consider this case as well. A process $\{Y_{id} :\ i = 1, 2, ...;\ 1 \le d \le D\}$ with unit Fréchet margins is max-stable if for any $n \ge 1,\ r \ge 1$,

$$\Pr\{Y_{id} \le ny_{id} :\ 1 \le i \le r;\ 1 \le d \le D\}^n = \Pr\{Y_{id} \le y_{id} :\ 1 \le i \le r;\ 1 \le d \le D\}. \tag{7.7}$$

Corresponding to this, a process $\{X_{id} :\ i = 1, 2, ...;\ 1 \le d \le D\}$ is said to be in the domain of attraction of a max-stable process $\{Y_{id} :\ i = 1, 2, ...;\ 1 \le d \le D\}$ if there exist normalising constants $a_{nid} > 0,\ b_{nid}$ such that for any finite r

$$\lim_{n\to\infty} \Pr\left\{\frac{X_{id} - b_{nid}}{a_{nid}} \le ny_{id} :\ 1 \le i \le r;\ 1 \le d \le D\right\}^n = \Pr\{Y_{id} \le y_{id} :\ 1 \le i \le r;\ 1 \le d \le D\}. \tag{7.8}$$

If we assume a priori that the X process also has unit Fréchet margins, then we may again take $a_{nid} = n,\ b_{nid} = 0$.

Smith and Weissman (1996) made a connection between max-stable processes and the limiting distributions of extreme values in dependent stochastic processes. In the one-dimensional case, again assuming for simplicity that the marginal distributions are unit Fréchet, there is a very large literature concerned with results of the form

$$\Pr\left\{\max_{1\le i\le n} X_i \le nx\right\} \to e^{-\theta/x} \tag{7.9}$$

where $\{X_i,\ i = 1, 2, ...,\}$ is a stationary stochastic process and θ is known as the *extremal index.* See for example Leadbetter *et al.* (1983), Leadbetter (1983), O'Brien (1987). If θ exists, it must be in the range $[0, 1]$ and is a constant for the process, i.e. it does not depend on x in (7.9).

For D-dimensional processes, the corresponding quantity is the *multivariate extremal index* defined by Nandagopalan (1990, 1994). For a stationary process $\{X_{id},\ d = 1, ..., D,\ i = 1, 2, ...,\}$ with unit Fréchet margins in each of the D components, we assume

$$\Pr\{X_{id} \le ny_d,\ 1 \le d \le D\}^n \to H(y_1, ..., y_D),$$

analogous to (7.1). The result corresponding to (7.9), for the joint distributions of sample maxima in the dependent process, is

$$\Pr\left\{\max_{1\le i\le n} X_{id} \le ny_d,\ 1\le d\le D\right\} \to H(y_1,...,y_D)^{\theta(y_1,...,y_D)}. \tag{7.10}$$

Equation (7.10), when it is true, defines the *multivariate extremal index* $\theta(y_1,...,y_D)$.

As with univariate stochastic processes, the multivariate extremal index satisfies $0\le\theta(y_1,...,y_D)\le 1$ for all $(y_1,...,y_D)$. Unlike the univariate case, it is not a constant for the whole process but it is true that $\theta(cy_1,...,cy_D)=\theta(y_1,...,y_D)$ for any $c>0$. Although it is by no means the only quantity of interest in studying extreme values of discrete-time stationary processes, because of its role in limit theorems such as (7.9) and (7.10), the (univariate or multivariate) extremal index is widely regarded as a key parameter.

The connection between max-stable processes and the multivariate extremal index is as follows. Suppose $\{X_{id}\}$ is a stationary D-dimensional process whose finite-dimensional distributions are in the domain of attraction of a max-stable process — in other words, there is a max-stable process $\{Y_{id}\}$ for which (7.8) is true for every $r\ge 1$. We also assume two technical conditions as follows:

I. For a given sequence of thresholds $u_n=(u_{nd}, 1\le d\le D)$ and for $1\le j\le k\le n$, let $\mathcal{B}_j^k(u_n)$ denote the σ-field generated by the events $\{X_{id}\le u_{nd},\ j\le i\le k\}$, and for each integer t let

$$\alpha_{n,t}=\sup\bigl\{|P(A\cap B)-P(A)P(B)| : A\in\mathcal{B}_1^k(u_n),\ B\in\mathcal{B}_{k+t}^n(u_n)\bigr\}$$

where the supremum is taken not only over all events A and B in their respective σ-fields but also over k such that $1\le k\le n-t$. Following Nandagopalan (1994), $\Delta(u_n)$ is said to hold if there exists a sequence $\{t_n,\ n\ge 1\}$ such that

$$t_n\to\infty,\ t_n/n\to 0,\ \alpha_{n,t_n}\to 0 \text{ as } n\to\infty.$$

II. Assume $\Delta(u_n)$ holds with respect to some sequence t_n, and define a sequence k_n so that as $n\to\infty$,

$$k_n\to\infty,\ k_nt_n/n\to 0,\ k_n\alpha_{n,t_n}\to 0.$$

We assume that, with $r_n=\lfloor n/k_n\rfloor$,

$$0=\lim_{r\to\infty}\lim_{n\to\infty}\sum_{i=r}^{r_n}\sum_{d=1}^{D}\Pr\left\{X_{id}>u_{nd}\middle|\max_d\left(\frac{X_{1d}}{u_{nd}}\right)>1\right\}. \tag{7.11}$$

Then we have:

Theorem 1. Suppose the processes $\mathbf{X}=\{X_{id}\}$ and $\mathbf{Y}=\{Y_{id}\}$ are each stationary with unit Fréchet margins, that $\mathbf{Y}$ is max-stable, and (7.8) holds. For a given $(y_1,...,y_D)$ and $u_{nd}=ny_d,\ d=1,...,D$, suppose $\Delta(u_n)$ and (7.11) hold for both $\mathbf{X}$ and $\mathbf{Y}$. Then the two processes $\mathbf{X}$ and $\mathbf{Y}$ have the same multivariate extremal index $\theta(y_1,...,y_D)$.

The practical implication of this is that in calculating the extremal index of $\mathbf{X}$, if suffices to look at the limiting max-stable process, assuming this exists.

7.3 Representations of Max-stable Processes

Suppose $\{\alpha_k,\ -\infty<k<\infty\}$ is a doubly-infinite sequence with $\alpha_k\ge 0$, $\sum_k\alpha_k=1$, and let $\{Z_i\}$ denote i.i.d. unit Fréchet random variables. Consider the *moving maximum* process

$$Y_i=\max_{-\infty<k<\infty}\alpha_kZ_{i-k}.$$

It is readily checked that Y_i is a stationary stochastic sequence with unit Fréchet margins, and moreover, is a max-stable process. The latter is most easily seen from the identity

$$\Pr\{Y_1 \le y_1, ..., Y_r \le y_r\} = \exp\left\{-\sum_{s=-\infty}^{\infty} \max_{1-s \le k \le r-s} \frac{\alpha_k}{y_{s+k}}\right\},$$

which satisfies (7.7) with $D = 1$.

Deheuvels (1983) defined a class of processes which, translated to the present context, we may call *maxima of moving maxima*, or M3, processes, as

$$Y_i = \max_{\ell \ge 1} \max_{-\infty < k < \infty} \alpha_{\ell,k} Z_{\ell,i-k},$$

where now $\{\alpha_{\ell,k}\}$ is a double sequence of constants satisfying $\alpha_{\ell,k} \ge 0$ and $\sum_\ell \sum_k \alpha_{\ell,k} = 1$, and $Z_{\ell,i}$ is a double sequence of independent unit Fréchet random variables. It is readily checked that this process is also max-stable.

Smith and Weissman (1996) generalised the M3 process to a multivariate context, the class of *multivariate maxima of moving maxima* processes, or M4 for short. In this definition, we assume a double sequence of independent unit Fréchet random variables $Z_{\ell,i}$ and a triple sequence of constants $\{\alpha_{\ell,k,d} :\ \ell \ge 1;\ -\infty < k < \infty,\ 1 \le d \le D\}$, such that $\alpha_{\ell,k,d} \ge 0$,

$$\sum_\ell \sum_k \alpha_{\ell,k,d} = 1 \text{ for each } d.$$

Then define the M4 process $\mathbf{Y} = \{Y_{id}\}$ by

$$Y_{id} = \max_\ell \max_k \alpha_{\ell,k,d} Z_{\ell,i-k}.$$

For this process, we have

$$\Pr\{Y_{id} \le y_{id},\ 1 \le i \le r,\ 1 \le d \le D\} = \exp\left\{-\sum_{\ell=1}^{\infty} \sum_{s=-\infty}^{\infty} \max_{1-s \le k \le r-s} \max_{1 \le d \le D} \frac{\alpha_{\ell,k,d}}{y_{s+k,d}}\right\}. \quad (7.12)$$

From (7.12), it is readily checked that $\mathbf{Y}$ is max-stable, and moreover, we can also show that the multivariate extremal index is given by the formula

$$\theta(y_1, ..., y_D) = \frac{\sum_\ell \max_k \max_d \alpha_{\ell,k,d} y_d}{\sum_\ell \sum_k \max_d \alpha_{\ell,k,d} y_d}. \quad (7.13)$$

The key characterisation result is the following: under some additional technical conditions which we shall not specify here, any D-dimensional stationary max-stable process with unit Fréchet margins may be approximated arbitrarily closely as the sum of an M4 and a deterministic process. This result was given by Smith and Weissman (1996) and is a direct generalisation of the result of Deheuvels (1983) for one-dimensional processes.

The deterministic process mentioned here is one that cycles infinitely through some finite set of values; since this seems unreasonable behaviour for most real observed processes, we usually assume it is absent. In that case, the practical interpretation of the result is that we can approximate any max-stable process with unit Fréchet margins by an M4 process.

Combined with the results of Section 7.2, this suggests the following strategy for analyzing extremal properties of multivariate stationary processes. First, fit the GPD or the point process model of Section 2.5 to the exceedances of the process above a threshold (possibly a different threshold for each component). Second, apply the transformation (7.3) so that the margins of the transformed process may be assumed to be

of unit Fréchet form above a threshold (7.4). Third, assume as an approximation that the joint distributions of the transformed process, above the relevant thresholds, are the same as those of an M4 process. Fourth, estimate the parameters of the M4 process $\{\alpha_{\ell,k,d}\}$. The multivariate extremal index of the process is then given by (7.13) and other extremal properties of the process may be calculated directly from the formula (7.12).

7.4 Estimation of Max-Stable Processes

So far, we have argued that max-stable processes form a suitable class of processes by which to study the extremal properties of D-dimensional stationary time series, and that after transforming the tail distributions to unit Fréchet margins, max-stable processes may be approximated by those in the M4 class. There are still some major difficulties, however, in estimating these processes in practice.

Since the M4 process is defined by an infinite set of parameters $\alpha_{\ell,k,d}$, to make progress we must in practice reduce these to a finite set. We therefore assume $\alpha_{\ell,k,d} = 0$ outside the range $1 \le \ell \le L,\ -k_1 \le k \le k_2$ where $L,\ k_1$ and k_2 are known. In the financial time series example to be discussed in Section 8, we have $D = 3$ and somewhat arbitrarily assume $L = 25,\ k_1 = k_2 = 2$.

Next, consider a situation where some value of $Z_{\ell,i}$, say Z_{ℓ^*,i^*}, is much larger than its neighbours $Z_{\ell,i}$ when $1 \le \ell \le L,\ i^* - k_1 \le i \le i^* + k_2$. Within this range, we will have

$$Y_{i,d} = \alpha_{\ell^*,i-i^*,d} Z_{\ell^*,i^*},\ i^* - k_1 \le i \le i^* + k_2,\ 1 \le d \le D.$$

Define

$$S_{i,d} = \frac{Y_{i,d}}{\max_{i^*-k_1 \le i \le i^*+k_2,\ 1 \le d \le D} Y_{i,d}},\ i^* - k_1 \le i \le i^* + k_2,\ 1 \le d \le D. \tag{7.14}$$

Then

$$S_{i,d} = \frac{\alpha_{\ell^*,i-i^*,d}}{\max_{i^*-k_1 \le i \le i^*+k_2,\ 1 \le d \le D} \alpha_{\ell^*,i-i^*,d}},\ i^* - k_1 \le i \le i^* + k_2,\ 1 \le d \le D. \tag{7.15}$$

The variables $\{S_{i,d}\}$ define a *signature pattern*, which specifies the shape of the process near its local maximum.

For any (ℓ^*, i^*), there is positive probability that (7.15) holds exactly. There are L such deterministic signature patterns (one defined by each ℓ^*), and when looking along the whole process for $-\infty < i < \infty$, it follows from elementary probability arguments that each of the L signature patterns will be observed infinitely often. Zhang (2002) has given some precise formulations and alternative versions of these results.

From an estimation point of view, this is both good and bad. On the one hand, if we were indeed to observe a process of exactly M4 form for a sufficiently long time period, we would be able to identify all L signature patterns exactly, and hence exactly estimate all the $\{\alpha_{\ell,k,d}\}$ parameters. However, in practice we would not expect to observe an exact M4 process, and even the theoretical results only go as far as justifying the M4 process as an approximation to the general max-stable process. Moreover, the presence of deterministic signature patterns means that the joint densities of the M4 processes contain singularities, which render the method of maximum likelihood unsuitable for such processes. Therefore, we must seek alternative methods of estimation.

Davis and Resnick (1989) developed many properties of "max-ARMA" processes (a special class of moving maxima processes) but did not devise a good statistical approach.

Hall, Peng and Yao (2002) proposed an alternative estimation scheme based on multivariate empirical distributions, for moving maxima processes. Zhang (2002) has generalised this to M4 processes.

It may also be possible to approach this question by assuming an unobserved process which is exactly of M4 form, and an observed process derived by adding noise, filtering the former from the latter by a Monte Carlo state-space approach. This has not so far been tried, however.

Here we propose a simple intuitive method for which we claim no optimality properties but which appears to be a useful practical approach. The steps of this new approach are as follows:

1. For each univariate series, fit the standard extreme value model to exceedances above a threshold and transform the margins to unit Fréchet via (7.3).

2. Fix the values of L, k_1 and k_2 such that $\alpha_{\ell,k,d} = 0$ except when $1 \leq \ell \leq L$, $-k_1 \leq k \leq k_2$. In our example we take $L = 25$ and $k_1 = k_2 = 2$.

3. For each local maximum above the threshold, define the signature pattern (7.14). In practice, some of the $Y_{i,d}$ values in (7.14) will be below their corresponding thresholds; in that case, set $S_{i,d} = 0$ for those values. In our example, $D = 3$ and there are 607 local maxima, so we end up with 607 candidate signature patterns in 15 dimensions.

Note that the theory specifies that L signature patterns occur infinitely often and these are the ones that identify the coefficients $\{\alpha_{\ell,k,d}\}$. In practice we will observe many more than L signature patterns, all different. The idea pursued here is to use a clustering algorithm to approximate the process by one in which there are only L signature patterns. Therefore:

4. Use the "K-means clustering" procedure to group the signature patterns into L clusters. K-means clustering is implemented in S-Plus and took less than a minute for the example given here.

5. Assume each cluster of signature patterns is represented by one signature pattern corresponding to the cluster mean, and estimate the $\alpha_{\ell,k,d}$ coefficients from the L cluster mean signature patterns.

In Section 8, we shall see the results of this procedure applied to financial data.

8. EXTREMES IN FINANCIAL TIME SERIES

The 1990s were mostly a boom time for the stock market, but some well-publicised catastrophes (e.g. Barings Bank, Orange County, Long-Term Capital Management) made investors aware of the dangers of sudden very sharp losses.

In response to this, there grew a new science of risk management. The best known tool is *Value at Risk* (VaR), defined as the value x which satisfies

$$\Pr\{X_T > x\} = \alpha,$$

where X_T is the cumulative loss over given time horizon T (e.g. 10 trading days), and α is a given probability (typically .05 or .01). However, X_T is typically calculated from a portfolio involving a large number of stocks, so the analysis of high-dimensional time series is involved.

Various tools have been devised to estimate VaR. The best known, but also the crudest, is *RiskMetrics*. Although this contains many complicated details, the main principles are:

1. The covariance matrix of daily returns on the stocks of interest is estimated from a fixed period (e.g. one month) of recent prices,
2. The distribution is assumed to be multivariate normal.

The purpose of the present discussion is to explore whether we could do better using extreme value theory, based on a specific example.

We use negative daily returns from closing prices of 1982-2001 stock prices in three companies, Pfizer, General Electric and Citibank. The data were plotted in Fig. 2.

Univariate extreme value models were fitted to threshold $u = .02$ with results given in Table 14.

Series	Number of exceedances	μ (SE)	$\log \psi$ (SE)	ξ (SE)
Pfizer	518	.0623	–4.082	.174
		(.0029)	(.132)	(.051)
GE	336	.0549	–4.139	.196
		(.0029)	(.143)	(.062)
Citibank	587	.0743	–3.876	.164
		(.0036)	(.119)	(.012)

Table 14. Parameters of the point process model of Section 2.5, fitted to each of the three financial time series based on threshold $u = .02$.

Diagnostic plots based on the Z and W statistics are shown for the Pfizer data in Fig. 37. The clear problem here is seen in the QQ plot for the Z statistics — this does not stay close to the assumed straight line, indicating that the point process of exceedance times is not well described by a uniform Poisson process. The reason for this is obvious to anyone familiar with the literature on financial time series: like all series of this nature, the Pfizer series goes through periods of high and low *volatility*, and extreme values are much more likely to occur in times of high volatility than low volatility.

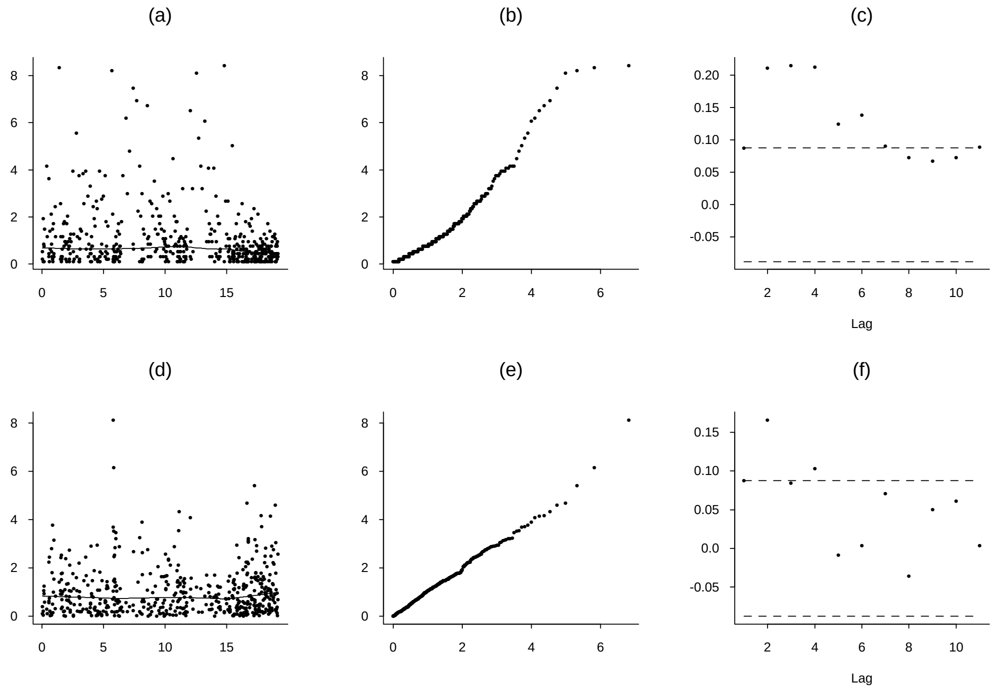


Figure 37. Diagnostic plots for the extreme value model fitted to the Pfizer daily returns.

For the present study, volatility was estimated by fitting a GARCH(1,1) model to each series. This is a familiar model in econometrics, see e.g. Shephard (1996). If y_t denotes the observed series (in this case, the observed daily return) on day t, assumed standardised to mean 0, then the model represents y_t in the form

$$y_t = \sigma_t \epsilon_t,$$

where $\{\epsilon_t\}$ are i.i.d. $N[0, 1]$ random variables, and the volatility σ_t is assumed to satisfy an equation of form

$$\sigma_t^2 = \alpha_0 + \alpha_1 y_{t-1}^2 + \beta_1 \sigma_{t-1}^2.$$

For the purpose of the present analysis, a GARCH(1,1) model was fitted to each of the three series, and a standardised series y_t/σ_t computed. The threshold analysis was then repeated for the standardised series, using threshold $u = 1.2$. Results from this analysis are shown in Table 15. The diagnostic plots are now satisfactory (plots for the Pfizer series are shown in Fig. 38; the others are similar).

Series	Number of exceedances	μ (SE)	$\log\psi$ (SE)	ξ (SE)
Pfizer	411	3.118	–.177	.200
		(.155)	(.148)	(.061)
GE	415	3.079	–.330	.108
		(.130)	(.128)	(.053)
Citibank	361	3.188	–0.118	.194
		(.157)	(.126)	(.050)

Table 15. Parameters of the point process model of Section 2.5, fitted to each of the three financial time series based on threshold 1.2, after standardising for volatility.

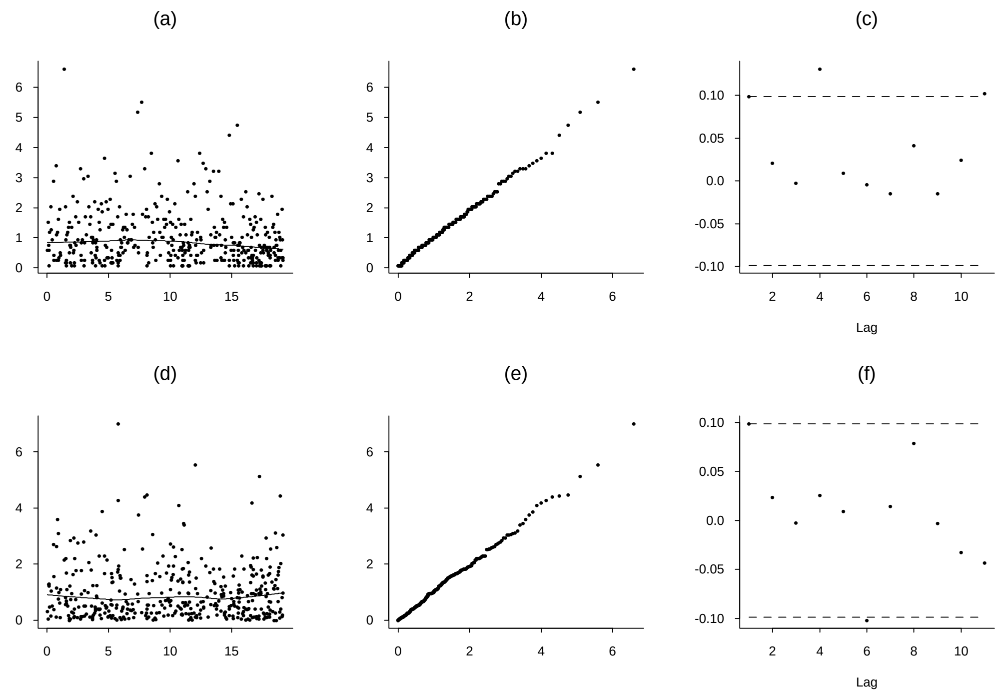


Figure 38. Diagnostic plots for the extreme value model fitted to the Pfizer daily returns, after standardising for volatility.

After fitting a univariate extreme value model to each series, the exceedances over the threshold for each series are transformed to have marginal Fréchet distributions. On the transformed scale, the data in each series consist of all values in excess of threshold 1. The resulting series are shown in Fig. 39.

Up to this point in the analysis, although we have taken into account long-range dependencies in the data through the volatility function, we have taken no account of possible short-term dependencies (e.g. serial correlation between values on one day and neighbouring days), nor have we considered any form of dependence among the three series.

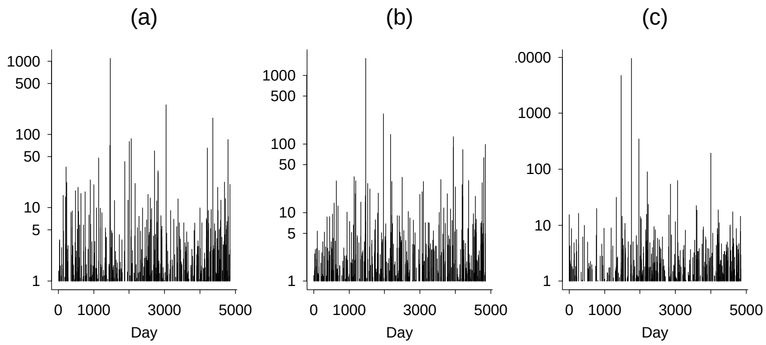


Figure 39. Plot of the threshold exceedances of standardised daily returns after transforming to unit Fréchet marginal distributions.

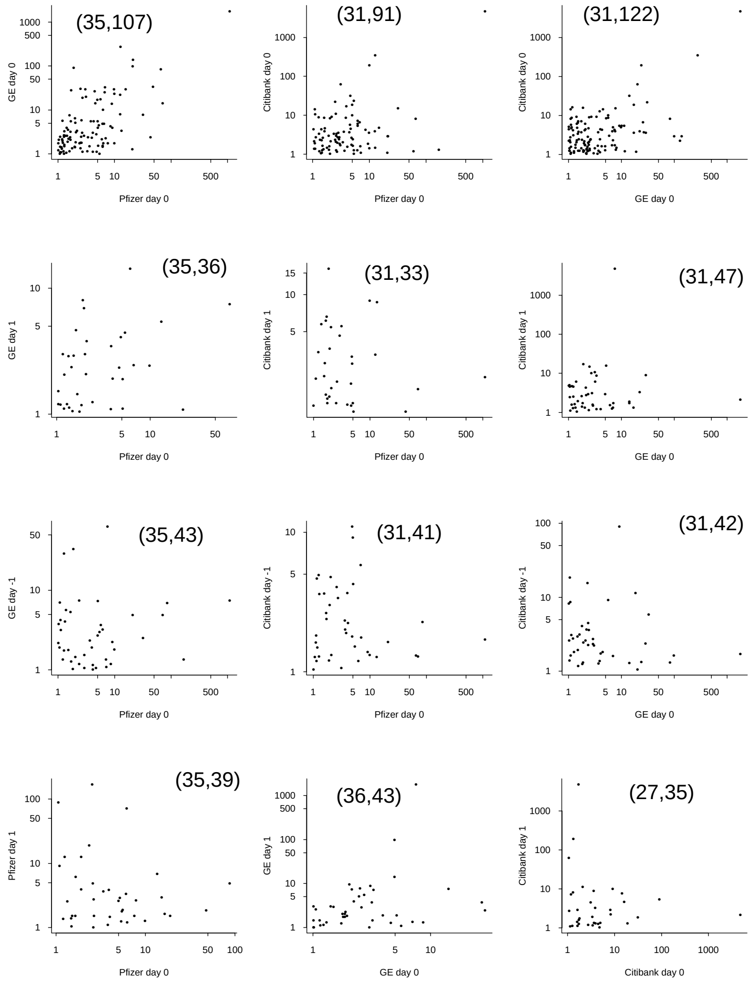


Figure 40. Scatterplots of the values over the threshold in the Fréchet transformed series. For each series, a scatterplot is shown against either the values of the other two series on the same day, or the values of all three series on either the day before or the day after. Of the two numbers shown on each plot, the second is the observed number of joint exceedances in the two series, i.e. the actual number of points in the scatterplot, while the first represents a calculation of the expected number of joint exceedances if the two series were independent (or for the case of one series plotted against itself, assuming the daily values are independent).

In Fig. 40, pairwise scatterplots are shown of the three transformed series against each other on the same day (top 3 plots), and against series on neighbouring days. The two numbers on each plot show the expected number of joint exceedances based on an independence assumption, and the observed number of joint exceedances.

Fig. 41 shows a plot of Fréchet exceedances for the three series on the same day, normalised to have total 1, plotted in barycentric coordinates. The three circles near the corner points P, G and C correspond to days for which that series alone had an exceedance.

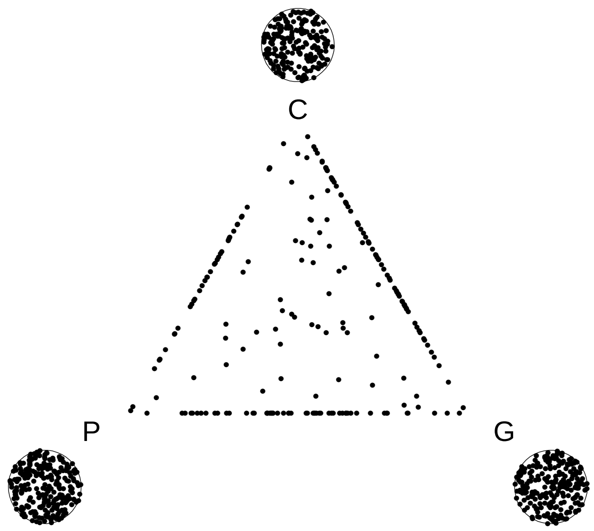


Figure 41. Representation of the trivariate distribution of threshold exceedances on the same day for the three Fréchet-transformed series. The letters P, G and C refer to the Pfizer, GE and Citibank series. The circle at each vertex of the triangle represent all the days that there was an exceedance in that series but neither of the other two series. For days on which at least two of the series had exceedances, the values were normalised to sum to 1, and plotted in barycentric coordinates.

These plots provide empirical evidence that there is dependence among the values of the three transformed series on the same day, and less clear evidence that there is also some serial correlation among the values on neighbouring days. These are the kinds of dependencies that the M4 series is intended to model, and the next step in the analysis is therefore to fit an M4 model to the threshold exceedances of the three transformed series. The method of fitting the M4 model follows the recipe given at the end of Section 7.

One test of whether the fitted M4 process is a realistic representation of the Fréchet-transformed time series is whether the sample paths simulated from the fitted process look similar to those from the original series. As a test of this, a Monte Carlo sample from the fitted model was generated, and Figs. 42–44 were drawn in the same way as Figs. 39–41 from the original series. One point to note here is that although the data were generated so that the marginal distributions were exactly unit Fréchet, in order to provide a fair comparison with the original estimation procedure, the marginal distributions were re-estimated, and transformed according to the estimated parameters, before drawing the scatterplots in Fig. 43 and 44. As a result, the signature patterns in the transformed simulated series are no longer the exact signature patterns of the M4 model, and the scatterplots are less clumpy than they would have been without the transformation. They are still more clumpy than the original Figs. 40 and 41. Despite this, we would argue that Figs. 42–44 provide a reasonable indication that the simulated series are similar to the original Fréchet-transformed time series.

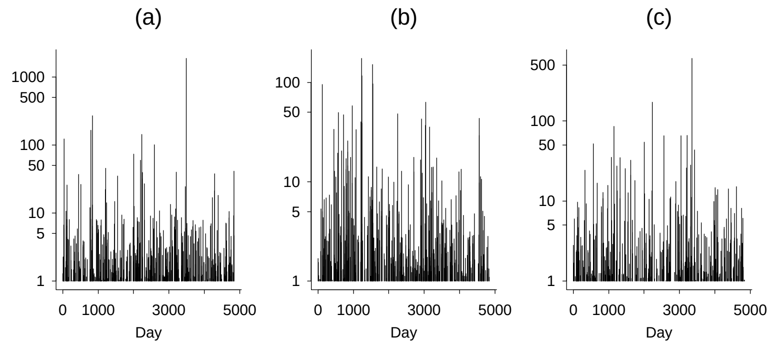


Figure 42. Same as Fig. 39, but with data simulated from the fitted M4 model.

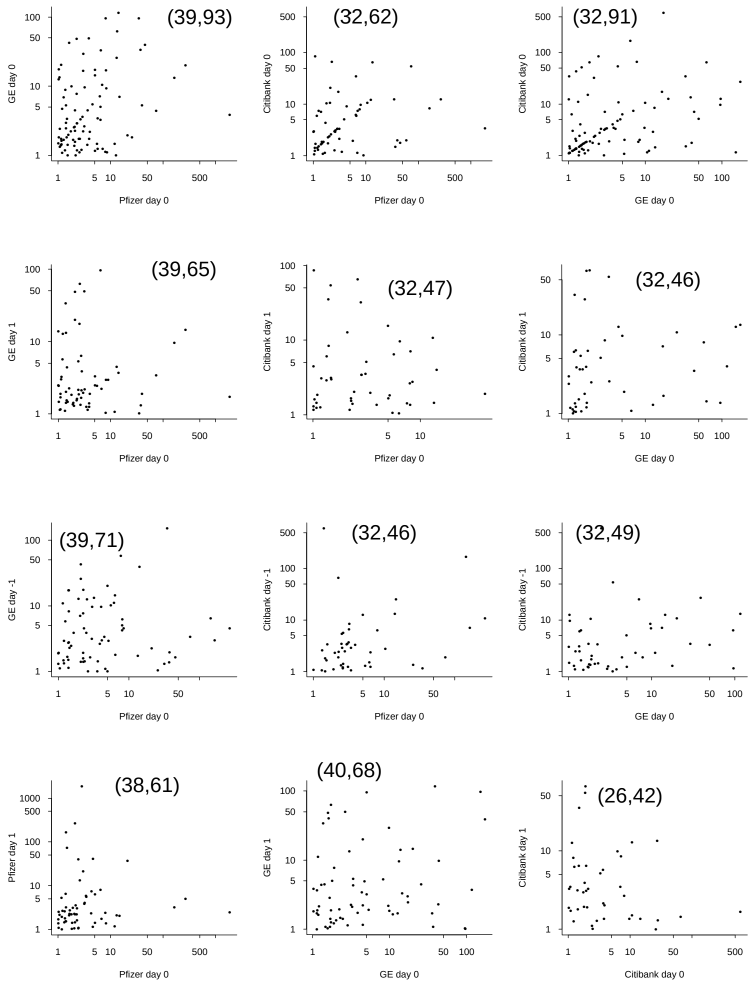


Figure 43. Same as Fig. 40, but calculated from the simulated data.

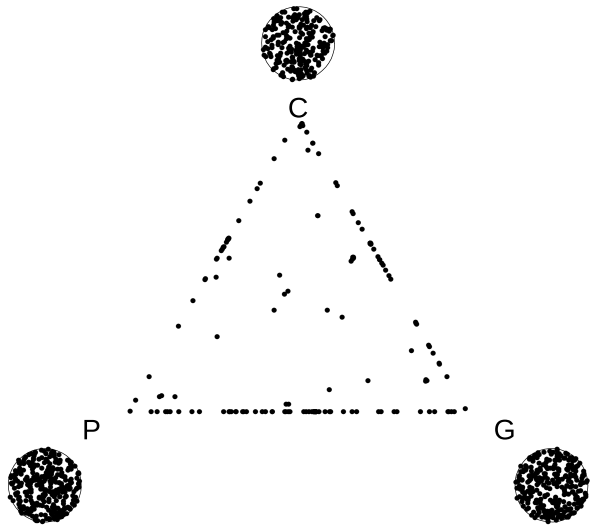


Figure 44. Same as Fig. 41, but calculated from the simulated data.

Finally, we attempt to validate the model by calibrating observed versus expected probabilities of extreme events under the model. The "extreme event" considered is that there is at least one exceedance of a specific threshold u, on the scale of the original daily return series, by one of the three series in one of the next 10 days after a given day. It is fairly straightforward to write down a theoretical expression for this, given the M4 model. To make the comparison honest, the period of study is divided into four periods each of length just under 5 years. The univariate and multivariate extreme value model is fitted to each of the first three 5-year periods, and used to predict extreme events in the following period.

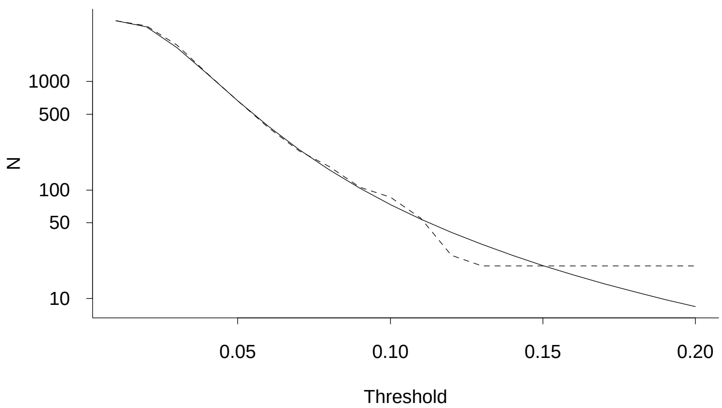


Figure 45. Representation of the probability $1/N$ that there is at least one threshold exceedance by one of the three series in a 10-day period, as a function of the threshold, on the scale of the original data. The solid curve represents the theoretical calculation based on the models fitted to the data. The dashed curve represents an empirical calculation of the same event, by counting observed exceedances.

Fig. 45 shows observed (dashed lines) and expected (solid lines) counts for a sequence of thresholds u. There is excellent agreement between the two curves except for the very highest thresholds, when the calculation may be expected to break down.

We conclude with some summary remarks.

The representation in terms of M4 processes contains the possibility of estimating both within-series and between-series dependence as part of the same model.

The key step in this method is the use of K-means clustering to identify a measure in a high-dimensional simplex of normalised exceedances. In contrast, existing methods of estimating multivariate extreme value distributions usually only work in low dimensions (up to 5 or so). However, this method of estimation is itself very *ad hoc* and the field is still open for more systematic methods of estimation with more clearly defined mathematical properties. The recent thesis of Zhang (2002) may point the way forward to some more general and powerful methods.

Ultimately the test of such methods will be whether they can be used for more reliable risk calculations than established methods such as RiskMetrics. The numerical example at the end shows that good progress has been made, but there are also many variations on the basic method which deserve to be explored.

9. REFERENCES

Cohen, J.P. (1982a), The penultimate form of approximation to normal extremes. *Adv. Appl. Prob.* **14**, 324–339.

Cohen, J.P. (1982b), Convergence rates for the ultimate and penultimate approximations in extreme value theory. *Adv. Appl. Prob.* **14**, 833–854.

Coles, S.G. (2001) An Introduction to Statistical Modeling of Extreme Values. Springer Verlag, New York.

Cressie, N. (1993), *Statistics for Spatial Data*. Second edition, John Wiley, New York.

Davis, R.A. and Resnick, S.I. (1989), Basic properties and prediction of max-ARMA processes. *Ann. Appl. Probab.* **21**, 781–803.

Davison, A.C. and Smith, R.L. (1990), Models for exceedances over high thresholds (with discussion). *J.R. Statist. Soc.*, **52**, 393-442.

Deheuvels, P. (1983). Point processes and multivariate extreme values. *J. Multivar. Anal.* **13**, 257–272.

Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997), *Modelling Extremal Events for Insurance and Finance*. Springer, New York.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications, Vol. I.* (3rd ed.) Wiley, New York.

Fisher, R.A. and Tippett, L.H.C. (1928), Limiting forms of the frequency distributions of the largest or smallest member of a sample. *Proc. Camb. Phil. Soc.* **24**, 180–190.

Fougères, A.-L. (2002) Multivariate extremes. This volume.

Gamerman, D. (1997), *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Texts in Statistical Science, Chapman and Hall/CRC Press.

Gnedenko, B.V. (1943), Sur la distribution limite du terme maximum d'une série aléatoire. *Ann. Math.* **44**, 423–453.

Grady, A. (2000), *A Higher Order Expansion for the Joint Density of the Sum and the Maximum with Applications to the Estimation of Climatological Trends*. Ph.D. dissertation, Department of Statistics, University of North Carolina, Chapel Hill.

Gumbel, E.J. (1958), *Statistics of Extremes*. Columbia University Press.

Hall, P., Peng, L. and Yao, Q. (2002), Moving-maximum models for extrema of time series. *Journal of Statistical Planning and Inference* **103**, 51–63.

Holland, D.M., De Oliveira, V., Cox, L.H. and Smith, R.L. (2000), Estimation of regional trends in sulfur dioxide over the eastern United States. *Environmetrics* **11**, 373–393.

Leadbetter, M.R. (1983), Extremes and local dependence in stationary sequences. *Z. Wahrsch. v. Geb.* **65**, 291-306.

Leadbetter, M.R., Lindgren, G. and Rootzén, H. (1983), *Extremes and Related Properties of Random Sequences and Series.* Springer Verlag, New York.

Nandagopalan, S. (1990). *Multivariate extremes and the estimation of the extremal index.* Ph.D. dissertation, Department of Statistics, University of North Carolina, Chapel Hill.

Nandagopalan, S. (1994). On the multivariate extremal index. *J. of Research, National Inst. of Standards and Technology* **99**, 543–550.

O'Brien, G.L. (1987), Extreme values for stationary and Markov sequences. *Ann. Probab.* **15**, 281-291.

Pickands, J. (1975), Statistical inference using extreme order statistics. *Ann. Statist.* **3**, 119-131.

Resnick, S. (1987), *Extreme Values, Point Processes and Regular Variation.* Springer Verlag, New York.

Robert, C.P. and Casella, G. (2000), *Monte Carlo Statistical Methods.* Springer Texts in Statistics, Springer Verlag, New York.

Robinson, M.E. and Tawn, J.A. (1995), Statistics for exceptional athletics records. *Applied Statistics* **44**, 499–511.

Shephard, N. (1996), Statistical aspects of ARCH and stochastic volatility. In *Time Series Models: In econometrics, finance and other fields.* Edited by D.R. Cox, D.V. Hinkley and O.E. Barndorff-Nielsen. Chapman and Hall, London, pp. 1–67.

Smith, R.L. (1985), Maximum likelihood estimation in a class of nonregular cases. *Biometrika* **72**, 67–90.

Smith, R.L. (1986), Extreme value theory based on the r largest annual events. *J. Hydrology* **86**, 27-43.

Smith, R.L. (1989), Extreme value analysis of environmental time series: An application to trend detection in ground-level ozone (with discussion). *Statistical Science* **4**, 367-393.

Smith, R.L. (1990), Extreme value theory. In *Handbook of Applicable Mathematics* **7**, ed. W. Ledermann, John Wiley, Chichester. Chapter 14, pp. 437-471.

Smith, R.L. (1997), Statistics for exceptional athletics records: Letter to the editor. *Applied Statistics* **46**, 123–127.

Smith, R.L. (1999), Trends in rainfall extremes. Preprint, University of North Carolina.

Smith, R.L. and Goodman, D. (2000), Bayesian risk analysis. Chapter 17 of *Extremes and Integrated Risk Management*, edited by P. Embrechts. Risk Books, London, 235–251.

Smith, R.L. and Shively, T.S. (1995), A point process approach to modeling trends in tropospheric ozone *Atmospheric Environment* **29**, 3489–3499.

Smith, R.L. and Weissman, I. (1996), Characterization and estimation of the multivariate extremal index. Preprint, under revision.

Tawn, J.A. (1988), An extreme value theory model for dependent observations. *J. Hydrology* **101**, 227-250.

Zhang, Z. (2002), *Multivariate Extremes, Max-Stable Process Estimation and Dynamic Financial Modeling.* Ph.D. dissertation, Department of Statistics, University of North Carolina, Chapel Hill.